# νόησις

# NOĒSIS

University of Toronto
Undergraduate Journal
of
Philosophy

Volume XVII

Spring MMXVI

# NOĒSIS

(Greek: intelligence, thought, understanding, mind). In Greek philosophy, the knowledge that results from the operations of *nous* (the mind, reason, intellectual faculty).

# NOĒSIS

UNIVERSITY OF TORONTO
UNDERGRADUATE JOURNAL OF PHILOSOPHY

# XVII



*Spring 2016*

# NOĒSIS

## UNDERGRADUATE JOURNAL OF PHILOSOPHY AT THE UNIVERSITY OF TORONTO

# XVII

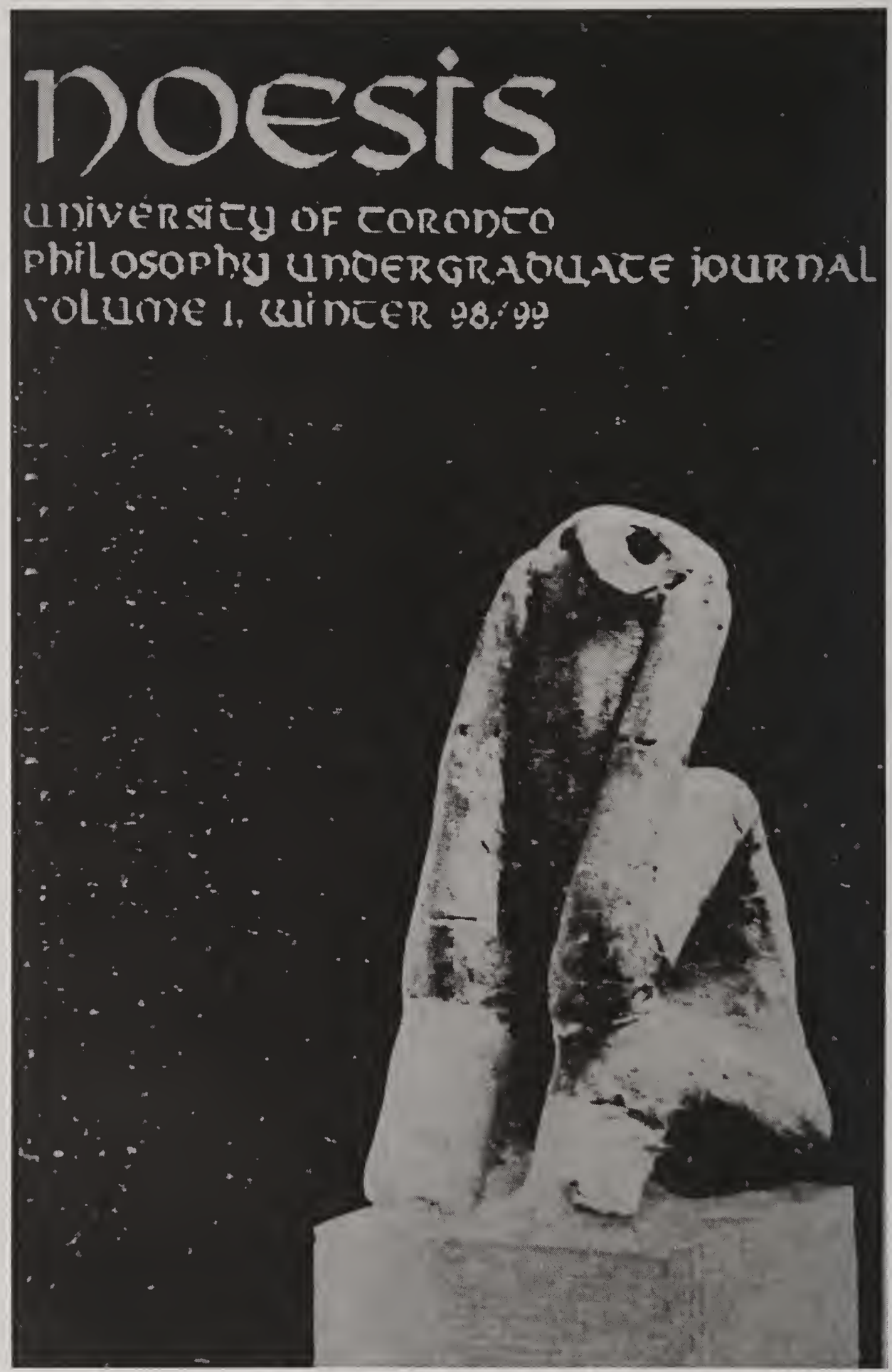
PAPERS



INTERVIEWS



NOĒSIS: University of Toronto Undergraduate Journal of Philosophy. Issue 17, Spring 2016.




**Above**: The cover of the first issue of Noēsis. The cover, designed by Ramona Ilea, features the famous Cucuteni-Trypillian statue "The Thinker of Tarpesti" (Gânditorul din Târpeşti). The statue dates around 4,500BCE, and is our oldest known "thinker". The Thinker of Tarpesti adorned the covers of Noēsis I–V.

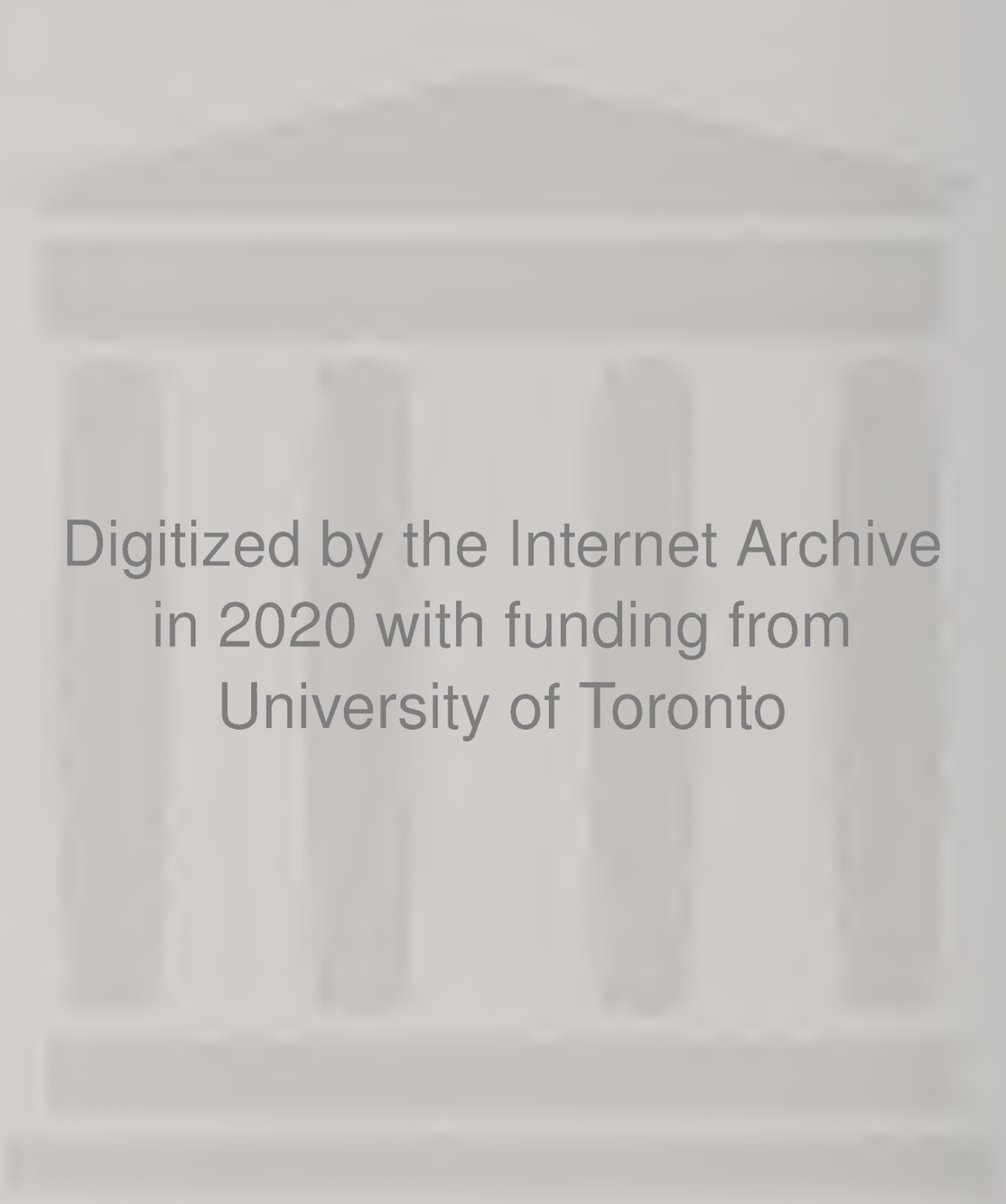
# Editors' Note

Welcome to Noēsis, the University of Toronto's undergraduate journal of philosophy. The aim of the journal is to provide a forum for philosophical work and discourse, and to foster a culture of lively philosophical exchange within the University of Toronto's undergraduate community. Noēsis is an undergraduate publication: from the original articles to the peer review process, editing and production, every aspect of the journal is the result of work done by our own undergraduates.

There are many individuals to whom Noēsis is greatly indebted for their continued support, and without whom the journal would likely cease to exist. We'd like to especially thank the officers of the Philosophy Department of the University for their continuing support of the journal and guidance for its editors. We also owe our gratitude to the members of the philosophy faculty for encouraging talented students to submit papers for publication, and to those many students who submitted to this issue of the journal. Finally, we owe our thanks to those professors who generously shared their time and thoughts with us for our interviews in this issue: Professors Imogen Dickie and Jon Rick.

This issue of Noēsis was made possible by the generous financial support of the Arts and Science Students' Union, the Department of Philosophy, Innis College, St. Michael's College, Trinity College, and University College. Thank you.

Noēsis would like to thank Alex CK Lui for his invaluable, stylish work in creating the design for the front and back covers of this issue. Well done, and thank you.



# Rejecting Jackson's Knowledge Argument with an Account of *a priori* Physicalism

Reggie Mills

## I. Introduction

In 1982 Frank Jackson presented the Knowledge Argument against physicalism: "Mary is a brilliant scientist who is, for whatever reason, forced to investigate the world from a black and white room *via* a black and white television monitor. She . . . acquires, let us suppose, all the physical information there is to obtain about what goes on when we see ripe tomatoes, or the sky, and use terms like 'red', 'blue', and so on."[1] Jackson asks, "What will happen when Mary is released from her black and white room or is given a colour television monitor? Will she *learn* anything or not?"[2] Jackson goes on to conclude, "It seems just obvious that she will learn something about the world and our visual experience of it."[3] Thus, something was missing from Mary's supposedly all-encompassing knowledge of the physical world—viz., an understanding of qualia. So, because complete physical information was insufficient for Mary's understanding of qualia, Jackson concludes that physicalism is false.

In this essay my objective will be to show that it is plausible for Mary to be able to know what it's like to have phenomenal experience while she's in the black-and-white room. I will start with some background on physicalism and present some responses Daniel Dennett has made to the Knowledge Argument. Then, I will outline the form of an *a priori* deduction of phenomenal truths from physical facts. Finally, I will show that at least some phenomenal concepts can be *a priori* deduced from lower-level properties, independent of experience. My hope is that by the end of the essay I will have made the case for *a priori* physicalism

---

[1] Frank Jackson, "Epiphenomenal qualia," *Philosophical Quarterly* 32 (1982): 130.
[2] *Ibid.*, 130. Emphasis Jackson's.
[3] *Ibid.*, 130.

stronger.

## II. Background on Physicalism

A definition of physicalism sufficient for our purposes is that there is nothing in the world except for what is specified by P, where P is a complete description of the world in the language of physics.[4,5,6] Since physics as we currently know it is incomplete, physics under this definition would refer not to our current knowledge but to an ideal/complete physics not radically different from our own in which all physical properties and truths are known.

Physicalism is widely (though not unanimously) agreed to be a contingent fact about our world.[7,8] In other words, it is possible for worlds to exist that aren't entirely specified by P. The main reason we are led to believe that physicalism is true is causal closure: We have no way to explain interactions between physical and nonphysical objects.[9,10] Although physicalism is contingent, a world in which physicalism is true is one in which there is a necessary metaphysical connection between P and the truths that P specifies. Of interest to us are phenomenal truths (Q = all phenomenal truths). To say that P does not necessarily give rise to the truths in Q would mean that something besides P determines Q, which is inconsistent with physicalism. Further, to say that Q can be nonphysical is also inconsistent. That qualia are nonphysical is what Jackson has concluded with the Knowledge Argument (KA).

The intuition that Mary learns something upon her experience of colour is convincing, I think, because there is an ineffable aspect to phenomenal knowledge; most people with properly functioning colour-

---

[4] Frank Jackson, *From Metaphysics to Ethics: A Defence of Conceptual Analysis.* (Oxford: Oxford University Press, 1998).

[5] David J. Chalmers, and Frank Jackson, "Conceptual analysis and reductive explanation." *The Philosophical Review* 110, no. 3 (2001).

[6] Robert Kirk, *The Conceptual Link from Physical to Mental.* (Oxford: Oxford University Press, 2013)

[7] Frank Jackson, "The Case for *a priori* Physicalism," in *Philosophy-Science-Scientific Philosophy, Main Lectures and Colloquia of Gap 5, Fifth International Confress of the Society for Analytical Philosophy*, edited by Christian Nimtz and Angsar Beckermann (2005).

[8] Robert Kirk, *The Conceptual Link from Physical to Mental.*

[9] W.V.O. Quine, *Theories and Things.* (Cambridge: Harvard University Press, 1981).

[10] Robert Kirk, *The Conceptual Link from Physical to Mental.*

vision grasp what it's like (w.i.l.) to have a phenomenal experience of red ($red_{ph}$). Yet, it seems nearly impossible to provide a reductive explanation for $red_{ph}$. That said, it is a stretch to conclude from the KA's intuition that physicalism is false. Mary is unable to deduce w.i.l. to experience colour from P. This is an epistemic issue concerning what is knowable; it lays no claim on the metaphysical connections that make up the world. I will now outline why this is the case.

Although Q is necessarily true given P, there are two ways that Q is knowable. The first is *a priori* physicalism, which holds that Q is knowable from P independent of any additional empirical information. The second is *a posteriori* physicalism, which holds that there are some truths in Q that cannot be deduced from P. It is not inconsistent for Q to be knowable only by experience despite the necessary P–Q connection. The classic example to show what is meant by a necessary *a posteriori* truth is the $H_2O$–water identity: it is not an *a priori* truth that water is $H_2O$ in all possible worlds; however, once it's discovered that water is $H_2O$ on a particular world, anyone with a grasp of the concepts "$H_2O$" and "water" will be able to realize the identity's necessity.[11] So, the two possibilities for P–Q's knowability are: (1) that all phenomenal truths Q are necessary *a priori* entailed by P; and (2) that all phenomenal truths Q are necessary *a posteriori* entailed by P.

The physical properties used need not be strictly micro-level and may be things like neural organization, as long as these properties are definable in physical terms. Since in (1) you are going from lower-level physical properties up to a grasp of w.i.l. to have phenomenal experience, I will refer to an *a priori*-type deduction as "bottom-up". And, since in (2) a grasp of w.i.l. to have a phenomenal experience is gained through having the experience (experience seeming like a higher-level phenomenon relative to physical properties), I will refer to learning knowledge *a posteriori* as "top-down". Necessary *a posteriori* physicalism would explain why phenomenal knowledge Q seems ineffable and nonreductive. If phenomenal properties in Q are knowable only through experience, there would not be a logical, epistemic connection between P and Q, despite a necessary metaphysical connection – no amount of information

[11] David Chalmers and Frank Jackson, "Conceptual analysis and reductive explanation."

from P would entail Q.

Since Mary is not able to deduce phenomenal knowledge from the physical facts, the KA is compatible with necessary *a posteriori* physicalism, in which phenomenal concepts can only be learned from a top-down approach. Since, given P, phenomenal truths Q are still necessarily true, qualia are not "left out of the physicalist story", as Jackson says.[12] Thus, physicalism is not challenged by the KA.

**III. Responses to the Knowledge Argument**

There have been attempts to reject the intuition that Mary would learn something when she leaves the room. In *Consciousness Explained* (1991), Daniel Dennett highlights that Mary's knowing all physical information (i.e., all of P) is a prospect so immense that it is nearly unimaginable for us; as such, it is difficult for those thinking about Mary to realize what she is capable of, and for this reason Dennett calls the KA an "intuition pump".[13] He goes on to propose what Mary's response should actually be when she sees colour, according to her knowledge of P. In his scenario, Mary's captors try to trick her by presenting her with a banana that's bright blue instead of yellow. Instantly, Mary points out the trick: She knew what physical impression a yellow banana was supposed to have on her nervous system and the thoughts that would accordingly result.[14] These thoughts, presumably, would entail for Mary an understanding of w.i.l. to see yellow. The implication then is that Mary has bottom-up deduced from P w.i.l. to see yellow without ever having phenomenally experienced yellow.

More recently, Dennett has come up with RoboMary, another Mary-like scenario to help us imagine Mary's capabilities and weaken the KA's intuitions.[15] RoboMary is a Mark 19 robot with the same complete knowledge of P that Mary has. She is largely identical to other Mark 19 robots in that her mental system is capable of processing visual information and giving her colour qualia, but she has one difference: Her

---

[12] Frank Jackson, "Epiphenomenal qualia," 131.
[13] Daniel Dennett, *Consciousness Explained*, (New York: Little, Brown, 1991), 398.
[14] Daniel Dennett, *Consciousness Explained*, 399.
[15] Daniel Dennett, *What RoboMary Knows,* ed. Torin Alter and Sven Walter. (Oxford: Oxford University Press, 2007).

visual sensors—her robot-substitute for human eyes—can detect only black and white. However, the scenario goes, RoboMary studies Mark 19s with functioning, colour-capable visual sensors and the processes that give rise to their colour qualia.[16] Using this knowledge, RoboMary is able to create a prosthesis that colourizes her black-and-white visual inputs to deliver herself colour qualia.[17] I think Dennett's intended implication is that RoboMary, in studying Mark 19 colourizing processes and then creating a likeness of such processes, *understands* the connection between lower-level physical properties and phenomenal properties.

To ensure that it is clear that RoboMary is not learning w.i.l. to see colour *a posteriori* from an experience of colour, Dennett further envisions Locked RoboMary, whose colour-experience registers—the systems in the Mark 19 brain that allow colour qualia to be experienced—are locked to greyscale. So, RoboMary cannot now experience colour qualia at all—neither through visual stimulus nor imagination. But, Dennett goes on, using some free RAM in her brain RoboMary constructs a simulated model of the Mark 19 visual processing system and uses it to calculate the mental states that would normally result from the phenomenal experience of coloured objects. Dennett refers to the nonphenomenal mental state after a phenomenal experience as a "dispositional state."[18] While this dispositional state is not itself an understanding of w.i.l. to have a certain experience, it contains all the information from such an experience that would be necessary for an understanding of w.i.l. So, comparing these colour-capable Mark 19s' dispositional states to those from her own black-and-white phenomenal experience, RoboMary makes it so that after her visual experience, she gets put into dispositional states as if she were a colour-sensing Mark 19. In other words, Locked RoboMary has calculated dispositional states containing understandings of w.i.l. to have certain colour experiences. I think we are to assume here that RoboMary again understands the P–Q connection, though this time without ever having experienced actual colour qualia.

---

[16] *Ibid.*
[17] *Ibid.*
[18] *Ibid,* 24.

RoboMary has received criticism from Torin Alter.[19] The way Alter reads RoboMary is that however RoboMary puts herself into her dispositional states, whatever goes on during the "putting" step itself does not confer an understanding of the P–Q connection.[20] RoboMary discovers the relevant colour-capable dispositional states, then "comes by her phenomenal knowledge of color experience not by *a priori* deduction from physical information but rather by putting herself in a nonphenomenal dispositional state that contains the relevant phenomenal information."[21]

I sympathize with Alter's view—here is how Dennett's RoboMary scenarios make sense for me: In the first, RoboMary copies the Mark 19 computational architecture to create a prosthesis which is able to generate colour qualia from black-and-white physical-level inputs. It does not seem that RoboMary would need to understand the processes going on in the Mark 19s to be able to copy the mental structure. So, the prosthesis does the hard work of transitioning from P to Q for RoboMary; all RoboMary is conferred by the prosthesis are its outputs—colour qualia. Thus, RoboMary gains an understanding of w.i.l. to have colour experience from colour qualia and not from P, in the same way that Mary in the KA gains an understanding of w.i.l. to experience colour after leaving the room.

In the second, Locked RoboMary builds a simulation which can take her black-and-white inputs, figure out their colourized equivalent, and confer to RoboMary the dispositional brain states that normally occur after *experiences* of these colours. The information that RoboMary is using to arrive at Q is not P, but rather information about phenomenal experience. So, in both cases, Q is arrived at by top-down means and not from a grasp of P, which is no better than Mary's arrival at Q in the KA.

I hold that to reject the KA, one must show that Mary is able to *a priori* deduce Q from P (i.e., a bottom-up deduction), rather than to arrive at Q from experience (i.e., top-down). Indeed, Jackson has referred to the specifics of such a bottom-up *a priori* deduction as "*the* hard issue that

---

[19] Torin Alter, "Phenomenal Knowledge without Experience," in *The Case for Qualia,* ed. Edmond Wright (2008), 247-267.

[20] *Ibid.*

[21] *Ibid.,* 253.

faces physicalists today."[22] Virtually all knowledge of Q and understanding of w.i.l. to experience colour for humans is learned top-down. But, Mary does not have direct access to colour experience while in the black-and-white room. In the remainder of this essay, I will shed some light on and argue for the plausibility of a bottom-up *a priori* P–Q deduction.

## IV. How *a priori* Physicalism Would Work

To be able to *a priori* deduce the truth of the conditional P ⊃ Q, one would need (a) sufficient empirical information from P such that the information implies a phenomenal concept in Q, plus (b) an understanding of the phenomenal concept in Q.[23] For example, regarding the water–$H_2O$ identity, "if a subject possesses the concept 'water' . . . then sufficient information about the distribution, behaviour, and appearance of clusters of $H_2O$ molecules enables the subject to know that water is $H_2O$, to know where water is and is not, and so on."[24] In the same vein, if P implies phenomenal truths in Q, then, given possession of a phenomenal concept such as $red_{ph}$ and sufficient empirical information from P, Mary plausibly could deduce the phenomenal truth that "This mug looks red." The empirical information in question would include things like the ~700-nm-wavelength photons emitted by the mug's surface, the detection of said photons by Mary's red-detecting opsins and the resulting electrical signal, and, importantly, the neurological organization in the human brain which induces the phenomenal experience of redness. To deduce the connection between P and $red_{ph}$ Mary would need to have a full grasp of all the necessary properties in P such that her grasp of these properties is equivalent to the concept $red_{ph}$.

Notably, none of this *a priori* deduction from P to Q involves Mary's experience of red; Mary needs only the empirical information from P and a grasp of the concept $red_{ph}$. Thus, the deduction is *a priori*/bottom-up. I do not know the specifics of the P–Q deduction, so the "hard issue" still remains. But, given that the connection from P to Q is metaphysically

---

[22] Frank Jackson, "The Case for *a priori* Physicalism, 264.
[23] David J. Chalmers and Frank Jackson, "Conceptual Analysis and reductive explanation."
[24] *Ibid.*, 323.

necessary, P–Q entailment seems plausible.

It is key to realize that the amount of information from P that Mary will need in order to have a grasp of truths in Q will not be trivial. Of particular note would be the neurological organization that gives rise to phenomenal experience and to the possession of phenomenal concepts. Whatever goes on here is something we evidently do not currently understand. Part of the reason for this, it seems, is that we just do not have enough empirical information about the brain to put together a cohesive explanation of consciousness; what we know from P is not enough to imply Q. For example, for $H_2O$–water, it would be difficult to deduce water from properties of $H_2O$ molecules if we did not know about the fundamental forces involved. But, it is also possible that no human will ever be able to have a grasp of properties from P that would entail phenomenal truths in Q. As Dennett says regarding his blue banana example, Mary knowing w.i.l. to experience blue "wasn't easy. She deduced it, actually, in a 4,765-step proof."[25] However, human incomprehensibility does not limit the apriority of the P–Q entailment. "Apriority concerns what is knowable *in principle*."[26] Consider also the apriority of $H_2O$–water: Even with an understanding of the concept "water", any human would be hard-pressed to deduce truths relating to water from empirical information about $H_2O$ molecules. But, the $H_2O$–water identity is still entailed *a priori*, and is routinely calculated via computer simulations.

The *a priori* deduction of Q from P that I showed above was presented with Mary already possessing a phenomenal concept such as $red_{ph}$, but the question we are interested in is if Mary could deduce w.i.l. to experience red from P without any pre-existing grasp of $red_{ph}$. So, we just need to rephrase the *a priori* deduction such that Mary uses sufficient empirical information from P to arrive at an understanding which would be equivalent to a grasp of the phenomenal concepts involved. Just as with $H_2O$–water, there will be a point at which a macroscopic grasp of the micro properties involved (from P) will be equivalent to a grasp of the higher-level concept (water or $red_{ph}$). In this way, possession of the concept $red_{ph}$

---

[25] Daniel Dennett, *What RoboMary Knows*, 16.

[26] David Chalmers and Frank Jackson, "Conceptual analysis and reductive explanation," 334. Emphasis mine.

is not a prerequisite for a bottom-up arrival at an understanding of the concept.

**V. Phenomenal Knowledge without Experience**

Part of the difficulty, I think, in imagining an arrival at phenomenal concepts without experience is that virtually every phenomenal concept we possess has been gained *a posteriori*. Let me here clarify that a grasp of a phenomenal concept is exclusive from the phenomenal experience itself. Think of the phenomenal concept of w.i.l. to experience pain ($pain_{ph}$). The thought of $pain_{ph}$ is not tied to an experience of pain, despite $pain_{ph}$ containing a full grasp of w.i.l. to experience pain. Similarly, understandings of other phenomenal concepts do not contain phenomenal experience. For taste and smell concepts I think the parallel to pain is obvious. For visual and auditory concepts, though, it seems that thinking of the concepts are almost unconscious triggers for imagining these concepts' phenomenal experiences. But, it is still possible with effort to think of these types of concepts without imagining or experiencing them. So, to reiterate, possessing a phenomenal concept is sufficient for understanding w.i.l. to have the relevant phenomenal experience. Then, a deduction of a phenomenal concept from lower-level properties would be sufficient to understand w.i.l. to have the corresponding phenomenal experience; nothing more would be learned by having the relevant experience once a concept is possessed.

Now we need to show how such an *a priori* (bottom-up) deduction of phenomenal concepts from lower-level properties is possible—i.e., tackle the hard issue. Here is an example to show at least that certain phenomenal concepts are bottom-up deducible: Someone familiar with a quadratic equation in the form $y = x^2$ can easily understand the parabolic shape of the function's curve, how changing the function to $y = 2x^2$ will make the parabola narrower, etc. This understanding is bottom-up deduced; we can change the equation to something novel and not previously experienced without compromising the apriority of the deduction. The phenomenal concept of the curve's shape ($curve_{ph}$), I think, is a simpler case in the many examples of phenomenal concepts. As I mentioned earlier, the *a priori* deduction of many of these phenomenal concepts (ones like $red_{ph}$ and $pain_{ph}$) from P would be quite difficult. I at

least think that the *a priori* deduction of $curve_{ph}$ from a quadratic equation is evidence that when lower-level properties contain higher-level phenomenal properties, a grasp of the lower-level properties can be equivalent to (and hence entail) an understanding of w.i.l. to experience those higher-level phenomenal properties. It should also be noted that $curve_{ph}$ is deduced from properties in the language of mathematics, an entirely different language from that of phenomenal concepts. So, transitioning between the potentially different languages of physical properties in P and phenomenal properties in Q should not affect the apriority of a P–Q deduction. Thus, I think it sounds plausible that Mary could have a grasp of properties in P equivalent to phenomenal concepts in Q without having had the corresponding experience for any Q-concepts, and this sort of P–Q *a priori* deduction would not be lacking any of the knowledge necessary for Mary to understand w.i.l. to experience phenomenal concepts.

**VI. Conclusion**

As I have tried to defend, I think it is possible in principle to *a priori* deduce all phenomenal truths and concepts from P. So, when Mary leaves the room, phenomenal concepts and knowledge of w.i.l. to experience these concepts will already be familiar to her; she would gain no knowledge. I have shown that it is possible to *a priori*, bottom-up deduce the phenomenal concept of a quadratic curve from a quadratic equation. However, I have no idea how the *a priori* deduction of complex phenomenal concepts like $red_{ph}$ and $pain_{ph}$ from P would work, and I leave the "hard issue" of consciousness unsolved for these. My belief is that the difficulty here is a result of human limitations, not a reflection of the metaphysical relationship between the physical and mental. I see there to be two possibilities for the current limitation: (1) That our understanding of the phenomenal concepts involved is insufficient, and (2) that we lack some of the necessary empirical information in P. Based on our current knowledge of the neurological organization of the brain I think (2) is undeniably true. Regarding (1), I am unsure if concepts such as $red_{ph}$ as we understand them are sufficient for a P–Q deduction. And, (1) and (2) need not necessarily be mutually exclusive; a possession of the phenomenal concepts involved sufficient to make the *a priori* deduction

of Q from P would include a grasp of the appropriate empirical information from P.

**Works Cited**

Alter, Torin. "Phenomenal Knowledge without Experience." In *The Case for Qualia*, ed. Edmond Wright, (2008): 247–267.

Chalmers, David .J. and Jackson, Frank. "Conceptual analysis and reductive explanation." *The Philosophical Review*. 110, no. 3(2001):315-360.

Dennett, Daniel. *Consciousness Explained*. (New York: Little, Brown, 1991).

—— *What RoboMary Knows.* in *Phenomenal Concepts and Phenomenal Knowledge*, ed. Torin Alter and Sven Walter. (Oxford: Oxford University Press, 2007).

Jackson, Frank. "Epiphenomenal qualia." *Philosophical Quarterly*. 32(1982):127–136.

—— *From Metaphysics to Ethics: A Defence of Conceptual Analysis.* Oxford: Oxford University Press, 1998.

—— "The Case for *a priori* Physicalism." In *Philosophy-Science-Scientific Philosophy, Main Lectures and Colloquia of Gap 5, Fifth International Congress of the Society for Analytical Philosophy*, ed. Christian Nimtz and Angsar Beckermann. Mentis. (2005).

Kirk, Robert. *The Conceptual Link from Physical to Mental*. Oxford: Oxford University Press, 2013.

Quine, Willard van Orman. *Theories and Things*. London: Harvard University Press, 1981.

# Is a Verification Machine Really a Problem for the Verifiability Principle? A Vindication of Lycan's Scepticism

Gavin Lee

In *The Concept of Meaninglessness,* Edward Erwin raises a novel objection against the verificationist program. William G. Lycan revisits Erwin's objection in *Philosophy of Language*. While he is critical of Erwin's claim of a decisive victory, he is unable to account for where Erwin's objection goes wrong. In this paper, I will critically examine Lycan's version of Erwin's objection, and using Lycan's framework, establish two ways in which the objection fails. I will start with a discussion of the verifiability principle, direct confirmation, and indirect confirmation. I will then present Lycan's version of the objection and demonstrate two ways in which it fails. Lastly, I will analyse how Lycan's presentation of the objection differs from Erwin's and explain why the differences are important.

The target of Erwin's objection is the verificationist program. The verificationist theory of meaning hinges on the verifiability principle, which states that a non-analytic statement is meaningful if and only if, in principle, it can be shown to be true or false through experience. A statement's meaning is its method of verification; we come to know the meaning of a statement when we know which conditions make it true and which conditions make it false.[1] For a verificationist, if we cannot conceive of an empirical method of verifying the truth or falsity of a statement, then that statement is meaningless.[2] Erwin distinguishes between the need for direct verification and indirect verification. Direct verification would be confirmation provided by immediate experience. Indirect verification is less stringent, allowing the use of induction and

[1] Moritz Schlick, *Positivism and Realism*, 86.
[2] *Ibid.*, 87.

probabilistic confirmation.[3] As Erwin points out, the verificationist theory is not tenable if it is taken as strictly demanding direct confirmation.[4] Therefore, indirect confirmation must be deemed as sufficient to provide meaning.[5] With this in mind, Erwin's objection takes aim at indirect verifiability.

Lycan's interpretation of the verifiability principle differs from Erwin's interpretation. In addition to the basic framework Erwin established, Lycan extends the principle by adding that a statement is meaningful "if and only if its being true would make some difference to the course of our future experience."[6] This requirement comes from Moritz Schlick, who argued that a statement is meaningless if its truth or falsity does not affect our experience.[7] For example, imagine I posit the existence of *entity-Z.* If somebody were to ask me, "In what empirically observable way does *entity*-Z's existence or non-existence make a difference in our world?" and my response was, "None whatsoever," then my statement about *entity*-Z is meaningless. With direct confirmation, it appears that this requirement is not an additional requirement at all; it is a restatement of the verifiability principle's primary claim, that a statement must be confirmable through experience. If we are able to verify a statement through direct observation, then what we are verifying simply *is* the difference in our world that the statement is making.[8]

When considering indirect confirmation, however, the situation is less clear. In the case of directly unobservable or undetectable entities, such as those posited by physics, we cannot always observe the difference they make in the world. When one of our physical theories posits a

---

[3] By probabilistic confirmation, I mean *degree* of confirmation, or probability of certainty.

[4] If indirect confirmation is not permitted by the verificationist principle, then the majority of our scientific statements would be evaluated as meaningless. Since any knowledge that we obtained using induction would be meaningless, we would only be left with the formal sciences.

[5] See Edward Erwin, *The Concept of Meaninglessness,* 34–5. Erwin provides examples of how direct confirmation is too demanding. Verificationist thinkers like Rudolf Carnap and A. J. Ayer readily ceded this point.

[6] Lycan, 98.

[7] Shlick, 88–9.

[8] This can be represented as a bi-conditional, where if we have verifiability, we have a verifiable difference in the world, and vice versa.

hypothetical entity, for which it cannot provide any account of *what* it does, simply that it is *predicted* or *required* by the theory, in what way can we even indirectly satisfy the requirement of it affecting our sensible experience? It may, in principle, be possible, but its 'in principle' seems to differ from the 'in principle' type of possibility that is raised by a question like, "Is there a mountain on Mars that is taller than Mt. Everest?" In principle, if the hypothetical entity did not exist, would that affect our future experience in any way? It could be argued that its non-existence would be counted as disconfirming evidence against our belief in the physical theory that predicted its existence.

Lycan and Schlick disagree about whether this type of effect on our experience satisfies the requirement. Does a change in our beliefs regarding a theory count as a change in our future experience? Under Schlick's account, the answer would be no, since he explicitly states that it must make a verifiable difference in the *sensible* world.[9] Schlick, then, is referring to direct confirmation in his statement.[10] A change in our understanding of a physical theory would not have an impact on what we observe, and that is what is required for Schlick. The change would only occur to our internal beliefs.[11]

Lycan broadens Schlick's conception to admit indirect confirmation when he says that a statement merely has to "make some kind of difference to thought and to action."[12] This widening allows for changes in internal beliefs to satisfy the requirement. For Lycan, the answer would presumably be yes, then, since there would be a verifiable difference in how we *think* about the affected physical theory. For Schlick, the requirements of verifiability and difference in future experience always

---

[9] *Ibid.*

[10] There is one sense in which an indirect confirmation could depend on a sensibly different world. A sensibly different world could affect a direct confirmation that informed a particular indirect confirmation. However, this indirect confirmation's reliance on the sensibly different world is derivative and would not be sufficient for Schlick.

[11] An objection could feasibly be raised here on the account of theory-ladenness. While it is clear that the theories we believe can affect our observations, they only affect them insofar as how we consider them. They do not affect what we actually observe in sensation, which is what Schlick cares about.

[12] Lycan, 99.

run together, but for Lycan, they can come apart.[13] Since Lycan's objection aims at indirect confirmation, we can safely set aside Schlick's conception.[14]

In his response to Erwin, Lycan presents his version of the objection through a thought experiment. Imagine we stumble upon a mysterious machine. Whenever we code a punch card with a declarative statement and submit it to the machine, it does some processing and then lights up as either TRUE or FALSE. We independently investigate the declarative statements that we submit to the machine, and we find that the machine's answers are always correct. Consider, then, that we coded an arbitrary statement *S* onto a punch card, and submitted it to the machine. Even if *S* is gibberish, if the machine returned TRUE, then we would have an empirical experience that would act as indirect confirmation for statement *S*. Since *S* could be any statement, we can confirm any statement, and the verifiability principle is trivial.[15]

With the context sufficiently sketched out, I will now move on to evaluating Lycan's version of the objection. According to Lycan, the only extant criticism of Erwin's objection comes from Brian R. Clack, who is suspicious of its "science-fiction character."[16] I hope to establish a more substantive attack on the soundness of Lycan's version. I will present two avenues of attack: An argument regarding the machine's logical consistency, and an argument about whether its assertions actually qualify as meaningful under the verifiability principle.

The first way that Lycan could attack his version of the objection is to focus on the fact that the machine is a *marvellous predictor* and is always right. In order to have the machine's assertions trusted as

---

[13] This would be possible if we had a statement that we could indirectly confirm or disconfirm without its truth or falsity making an appreciable difference to our future experience. In this case the bi-conditional would fail, as verifiability would not entail a difference.

[14] As the machine's judgments are examples of indirect confirmation (they are trusted because of an inductive inference), it follows that the objection is attacking indirect confirmation.

[15] William G. Lycan, *Philosophy of Language,* 104. If all statements can be confirmed, then all statements are meaningful. If all statements are meaningful, and the verifiability principle is supposed to demarcate meaningful statements from meaningless ones, then it ceases to do any work.

[16] Brian R. Clack, *Religious Belief and the Disregard of Reality*, 270.

confirming evidence, the people in the thought experiment would need to first establish with indirect confirmation that it is, indeed, always right.[17] Something like the following would suffice:

1. Provide the machine with declarative statements, which they believe to be true, and have it respond TRUE.
2. Provide the machine with declarative statements, which they believe to be false, and have it respond FALSE.
3. Provide the machine with declarative statements, of whose truth or falsity they are uncertain, and have it respond TRUE or FALSE, and then after further investigation, confirm its findings.

The people in the thought experiment may also provide the machine with statements they believe to be true or false, have it respond to the contrary, and upon further investigation, they would confirm its findings. The machine does all of this perfectly, with no mistakes. This brings us to the example Lycan uses in the thought experiment, *ST4*:

4. Provide the machine with a statement of gibberish, which they believe to be meaningless, and have it respond TRUE.

They might then ask: why is the machine evaluating this statement as TRUE? The only reason the thought experiment provides is that the machine is *always right.* We can use this fact to deduce the following:

I. If the machine is always right, it must evaluate all types of statements accurately, whether they take the form of 1, 2, 3 or 4.
II. Since we know through *ST1-3* that the machine's evaluations accurately reflect our empirical investigations, we can conclude that its evaluations cannot be arbitrary, and that they must reflect the world.[18]
III. Therefore, the machine evaluates statements like *ST4* the same way it does for statements like *ST1-3,* by returning TRUE or FALSE.
IV. Therefore, when the machine returns TRUE or FALSE to

---

[17] This is similar to a famous mathematician having to prove several important theorems before we would regard her authorship in itself as confirming evidence.

[18] In *ST1-3* we see that the results of our empirical investigations in the world perfectly match the assertions made by the machine. If our assertions track the world, and the machine's track ours, then it follows via transitivity that the machine's assertions are tracking the world.

> statements like *ST4*, those gibberish statements must actually reflect the world in the same way that statements like *ST1-3* do.[19]

With this in mind, let us imagine that we continue to feed the machine statements like *ST4*. We continuously feed the machine with an indefinite succession of unique statements of gibberish. Each time the machine processes a new statement, if it returns TRUE or FALSE, then the statement must have been about something in the world, even if we cannot understand how.[20] If what we submit ever happens to be incomprehensible or does not pertain to the world, then the machine would not be able to respond with TRUE or FALSE. This makes intuitive sense, for how could the machine consistently evaluate statements like *ST4* as TRUE or FALSE if they were not connected to the world at all? In this case, the machine would respond with something like NULL. For the objection to succeed, it requires every arbitrary statement to be meaningful, which leads to the trivialization of the verificationist principle. Since the statements evaluated by the machine as NULL would not be meaningful, if the NULL response ever occurs, then the objection to the verificationist principle fails.

The only way in which the objection can be saved from this attack, then, is if every conceivable statement is both comprehensible and about something in the world.[21] Even if we grant that our knowledge is limited, and that some statements could be comprehensible and be about something in the world in a way we do not currently understand, it seems inconceivable that every possible statement we do not understand is like this. This intuition can be confirmed by the following. The set of 'conceivable statements,' *CON*, that could be used in statements of the

---

[19] Restating this negatively may make this clearer: if statements like *ST4* were not about the world, that is, if they were unlike in kind to statements like *ST1-3*, then the machine could not return TRUE or FALSE in response to them, as it does for *ST1-3*, as neither evaluation would be accurate. And the machine *must* be accurate.

[20] This is not a problem for the verifiability principle, as something's being verifiable *or* falsifiable is sufficient for meaning.

[21] If every statement can be meaningful, then for meaningfulness to be non-trivial (which the verificationists need to be the case) every statement must be about something in the world.

kind *ST4* is really the set of *all* possible statements.[22] However, each statement in the set of all possible statements is not a coherent statement about the world. Therefore, the set of 'coherent statements that can be made about the world,' *COH,* is a proper subset of *CON*.[23] Therefore, there must be statements in *CON* that are not in *COH,* and when one of these statements is submitted to the machine in the form of *ST4,* the machine would not return TRUE or FALSE, and the objection fails.

There are a few counterpoints to address here. First, someone could claim that the machine is a perfect predictor for statements like *ST1-3*, but not for statements like *ST4*. This is inconsistent with the foundational premise of the thought experiment, as it is stipulated that the machine is always correct. Second, and more interestingly, someone could claim that up until $t_1$, the point in which we first submit a statement like *ST4* to the machine, the machine is always right, and that after that point in time it becomes inconsistent. The people in the thought experiment would be able to establish the consistency of the machine pre-$t_1$ while avoiding the possibly of potential NULL responses post-$t_1$.[24] The problem here is that they would still continue to provide statements like *ST1-3* to the machine and they would still follow up on them from time to time. They would discover that the machine was returning incorrect results to statements like *ST1-3*, and then would come to no longer view its answers as providing confirmation. Once this occurs, the objection fails. This relates back to Erwin's point about renowned mathematicians.[25] If a famous mathematician published a long string of flawed work, we would no longer view the fact that it was her making an assertion as evidence for that assertion being true.

The second way Lycan could attack the objection stems from his own widening of the verifiability principle. Let us grant that it is possible

---

[22] Since the verificationists are not metaphysical realists about statements, statements are the kind of thing that can only be created by a mind. In order to create a statement, a mind would first need to conceive of it.

[23] B is a proper subset of A if and only if B is a subset of A and there exists at least one element of A which is not in B. In this case, $COH \subsetneq CON$.

[24] In this case, the machine could respond TRUE or FALSE to statements that would normally prompt a NULL response. Since the machine does not have to be always correct after $t_1$, wrong answers are tolerated.

[25] Erwin, 36.

for the machine to always be right and return TRUE or FALSE for any arbitrary *ST4*-like statement. Let us call such an arbitrary statement *S*. These TRUE or FALSE evaluations allow for any *S* to have an empirical verification condition, that is, the people in the thought experiment will see the machine respond TRUE or FALSE to their submitted *S*, and that acts as indirect confirmation for *S*. This satisfies the first statement of the verifiability principle, which simply states that the statement must be empirically verifiable.

For Lycan's second requirement, however, things are less clear. Does the truth or falsity of any *S* change the future experience of any of the people in the thought experiment? Does it make some kind of difference to their thoughts and actions? It does not appear that it would. The truth or falsity of any *S* does not influence anyone's beliefs about any theory, the world, or even the beliefs they hold about the machine itself. They cannot independently investigate any arbitrary *S* like they can for statements like *ST1-3*. Therefore, the recursive mechanism that enables the independent corroboration of statements like *ST1-3* to affect the people's beliefs about the machine's reliability is not available. The TRUE or FALSE evaluation of any *S* does not tell the people anything about the world which could impact their future experience. Lycan describes the difference requirement as a question: What will happen depending on whether the statement is true or false?[26] In this case, posed to any arbitrary *S*, the answer would be nothing.

One could challenge that an arbitrary statement *S*'s truth or falsity does make a difference, and that the difference simply is the machine responding with TRUE or FALSE.[27] This raises a question of priority. When Lycan says a statement's truth or falsity must make some difference to the course of future experience, I am reading 'future' to mean *after* a statement's truth or falsity is determined. This must be the correct reading, because to assert that the difference in *future* experience simply *is* the response, we would have to presuppose that its truth or falsity is already known to recipients before they receive it from the machine. This is not the case. Once they receive the TRUE or FALSE response, their future

---

[26] Lycan, 104.

[27] To be clear, the difference simply *is* receiving the response from the machine that *S* is TRUE or FALSE.

experience begins, and it continues without difference.[28] Even if we did grant that the response of TRUE or FALSE itself is a difference in future experience, then it would render Lycan's second requirement vacuous – there could be no true or false statement that did not satisfy it, since its very evaluation would qualify.

Someone could further object, "Well, surely the recipients will think about S's truth or falsity in some way after receiving the response from the machine, even ever so briefly, to discard the answer as useless, and that very thought satisfies Lycan's requirement." This falls short because Lycan stipulates that the truth or falsity must make a difference both for thought and for action, and just discarding S makes no difference for a person's future action.[29] This points to a bigger problem. This challenge takes Lycan to mean literally any thought when he says 'thought'. While Lycan does not make it explicitly clear what exactly he means by thought, a reasonable account should presume a more nuanced conception than the loosest one possible.

It is interesting that Erwin's version of the objection differs from Lycan's version. Erwin's version of the objection is presented in the form of a thought experiment as well. He asks us to imagine that there exists a computer that knows all the information there is to know. The nature of the computer is such that it only makes true assertions. Suppose we start to receive assertions made by the machine, and every time we independently attempt to confirm them, we find them to be true, and we can repeat the process indefinitely. After enough time has passed, if the machine were to make any assertion *A,* and we did not know whether *A* was true, false, or meaningless, the mere fact that the machine was asserting *A* would be sufficient to indirectly confirm *A*. As *A* could be any statement, it follows that for any conceivable statement made by the

---

[28] There is one sense in which this is not true. In the future, a statement of the kind, "At Time *X*, the machine read TRUE about statement *S*" will become true. However, as I explain, this type of difference is insufficient as it would render Lycan's second requirement vacuous.

[29] There is room for here for an objection. Instead of merely mentally discarding the result, the person could instead crumple up and throw away the printed judgment while exclaiming "This is useless!" Similar to the discussion that follows regarding the looseness of the term 'thought,' I think it is justified to take Lycan to be meaning some more robust sense of 'action' than what is exemplified by this example.

machine, we would have a method of verifying its truth. Therefore, any conceivable statement made by the machine would be meaningful. If *any* conceivable assertion can be meaningful, then the verifiability principle is trivial.[30] Erwin argues that if his objection is sound, then it proves fatal to the verifiability principle.

In isolation, the thought experiments differ in at least one important way. In Erwin's version, we do not actively interact with the machine. That is, we only receive assertions from the machine, and we cannot have it judge any assertion that we provide to it. Lycan's presentation of the thought experiment allows us to be active participants, with us submitting assertions for the machine to consider. My first attack on Lycan's version of the objection heavily relies on *ST1-3* and *ST4* in order to deduce I-IV. Since we cannot actively submit statements to the machine in Erwin's version, this same approach may not work.[31] For my second attack, Erwin's statement of the verifiability principle does not include anything about a requirement for a difference to be made in future experience. While further investigation would be required in order to determine whether Erwin's version is susceptible to my attacks, it seems likely that it may not be.

If my two criticisms of the objection are successful, and I believe that I have demonstrated that they are, then Lycan's scepticism is well-placed. A careful analysis of the machine's nature shows that the claimed results of the thought experiment would not occur. Further, even granting that the claimed results would occur, they would not satisfy Lycan's requirements for any arbitrary *S* being a meaningful statement. The result is that Lycan's presentation of the objection does not threaten the verificationist principle, which renders verificationism safe from this particular thought experiment. However, Lycan's scepticism is only well-

---

[30] *Ibid.,* 36–8. Erwin points to real-world examples in mathematics and physics in order to demonstrate how we use this type of induction. Mathematical and physical assertions made by prize-winning mathematicians and physicists are lent some measure of confirmation simply in virtue of the past success their proposers have had in their respective fields.

[31] The machine may never output statements like *ST4*, or it might only do so very sporadically. Without being able to systematically study the machine by submitting statements to it, we may not be justified in treating its random outputs as indirect confirmation.

placed as applied to his own version of the objection. Since it is not clear whether my attacks would be successful against Erwin's version of the objection, the verificationist principle may still require additional defense.

## Works Cited

Clack, Brian R. "Religious Belief and the Disregard of Reality." In Joseph Carlisle, James Carter, and Daniel Whistler, eds., *Moral Powers, Fragile Beliefs* (New York: Continuum International Publishing Group, 2011).

Erwin, Edward. *The Concept of Meaninglessness* (Baltimore: The John Hopkins Press, 1970).

Lycan, William G. "Verificationism." In *Philosophy of Language: a Contemporary Introduction* (New York: Routledge, 2000).

Schlick, Moritz. "Positivism and Realism" in A. J. Ayer, ed., *Logical Positivism* (New York: The Free Press, 1959).

# Does Sherlock Holmes Exist? A Criticism of van Inwagen's Theory of Fictional Objects

Lucas Bennett

## I. Introduction

Characters are very curious things indeed. Each one of us knows of literally hundreds of characters and can even provide a set of descriptions about each one. And yet, we know that not a single one of them exists. However, according to Peter van Inwagen this categorization is a complete mistake. He takes an artifactualist approach, where characters are not a sub-set of non-existent objects, but rather are abstract entities. In fact, not only do characters exist, according to van Inwagen, but so does every other fictional creation made by an author, including places (such as "Narnia") and things (such as "The One Ring"). Together, van Inwagen calls these fictional objects "creatures of fiction."[1,2] I argue that van Inwagen's model fails as a successful account of fictional objects since it rests upon certain faulty assumptions about reference and ordinary language, and it is insufficient to support his distinction between predication and ascription. However, as I aim to show, van Inwagen's failure can help guide us to the beginnings of a correct ontological theory of creatures of fiction.

## II. Setting Up the Problem

Given our ordinary talk of fictional objects, a common problem is how to make sense of seemingly meaningful referential statements about things that do not exist. Reference is a relation obtaining between a sort of representational token (such as a name, a picture, or an idea) and a certain object. For instance, if I give the statement,

---

[1] In what follows, I will use these terms interchangeably.

[2] Peter van Inwagen, "Creatures of Fiction," *Philosophical Quarterly* 14, no. 4 (1977): 302.

(1) Barack Obama is the President of the United States,

I am using the name "Barack Obama" as a representational token to identify or pick out a particular individual in the world and say something true about that individual, namely, that he is President of the United States.

The problem arises when we make similar referential statements about fictional objects. For instance, take the following statements:

(2) Sherlock Holmes is a detective who lives on 221B Baker Street.
(3) Sherlock Holmes is a milkman who lives in Saskatoon.

These statements, like (1), are in subject-predicate form (*S* is *p*), but while (1) and (2) both seem to be true statements, (3) seems to be false. Typically, we take it that a statement is true so long as it corresponds to reality. (1) is true because the name "Barack Obama" successfully refers to an object in the world and that object is actually the President of the United States. Obviously if Barack Obama was not the President but a senator, then (1) would be false.

But this strategy does not seem to work for (2) and (3) since, by hypothesis, the name "Sherlock Holmes" does not pick out a particular individual in the world, and so there is nothing in the corporeal world for the statements to correspond to. But then, how is it that we can intuitively identify (2) as being true and (3) as being false? This is what we may call the problem of reference:

(a) Fictional objects do not exist.
(b) There are true (and false) statements about fictional objects.
(c) If a statement of the form *S* is *p* is about *S*, then there exists an *x* such that *S* refers to *x*.

Although we commonly take these propositions to all be true, at least one of them must be false since they are logically inconsistent.

As it stands, we must clarify what we mean by "statements about fictional objects." I take it that there are at least three different types of statements about fictional objects, each of which may be subject to a

different analysis.[3] First, there are fictional assertives, which are descriptions made by authors about fictional objects.[4] Closely related to fictional assertives are literary descriptives, which are statements made by speakers other than the author and are about fictional objects that have already been written about. Finally, there are meta-fictional statements that describe relations between fictional objects and the real world. Examples of these are:

(i) "[Sherlock Holmes] was rather over six feet, and so excessively lean that he seemed to be considerably taller. His eyes were sharp and piercing…and his thin, hawk-like nose gave his whole expression an air of alertness and decision."[5] (Statement written by Doyle)
(ii) Sherlock Holmes was six feet tall, excessively lean, and had sharp piercing eyes and a thin hawk-like nose. (Statement given by a person other than Doyle after Doyle has written about Sherlock).
(iii) Sherlock Holmes is more famous than any detective today.

In the first half of this paper, I will examine the foundation of van Inwagen's solution to the problem of reference, and in the second half, I will address his analysis of fictional assertives, literary descriptives, and meta-fictional statements.

## III. The Quantification Argument for the Existence of Fictional Objects

The first step in van Inwagen's solution to the problem of reference is to reject proposition (a), the idea that fictional objects are really non-existent objects. There are some caveats to his theory, for while he maintains that creatures of fiction do exist, he grants that they do have a much different ontology than the regular objects that we are familiar with (chairs, trees,

---

[3] Though van Inwagen does not give an explicit taxonomy of statements about fictional objects, it is clear that he accepts something along these lines, see van Inwagen, "Creatures of Fiction," 301.

[4] Although this is what it appears authors are doing when they make statements about their fictional objects, I think that this is fundamentally incorrect, as I will argue in section IV.

[5] Arthur Conan Doyle, *A Study in Scarlet* (New York: Modern Classics Library, 2003), 12.

airplanes, planets and so forth).[6] Fictional objects belong to a broader ontological category of "theoretical entities of literary criticism."[7] Included in this category are not only characters, places, and objects mentioned in stories, but also rhyme schemes, literary forms, plots, and so on, all of which have an abstract as opposed to a concrete existence.[8]

Take the following meta-fictional sentence given by van Inwagen:

(4) "There are characters in some 19th–century novels who are presented with a greater wealth of physical detail than is any character in any 18th–century novel."[9]

Meta-fictional statements like this are quite common in literary criticism, and they all seem to assert the existence of characters.[10] After all, to give a statement like (4), it would seem that it could only be true so long as there were such things as characters in novels. If there were no such things, what could a statement like (4) be about? To use Quine's terminology, it seems that we are "ontologically committed" to the existence of characters from sentences like (4).[11]

This fact is even more apparent when we apply formal logic. Rendering (4) in quantifier idiom, we get

(4*) $\exists x\,(C(x) \,\&\, \forall y\,(N(y) \rightarrow P(x, y)))$

Where $C(x)$ is "$x$ is a character in a 19th century novel," $N(y)$ is "$y$ is a character in an 18th century novel," and $P(x, y)$ is the two place predicate, "$x$ is presented with a greater wealth of physical detail than is $y$."
But, by the rules of formal logic, we may derive the following:

(5) $\exists x\, C(x)$

This simply means that there is some $x$ such that $x$ is a character in a 19th-century novel. Hence, there must exist characters, a fact which we appear

---

[6] van Inwagen, "Creatures of Fiction," 303.
[7] Ibid.
[8] Ibid.
[9] Ibid., 302.
[10] Along with van Inwagen, I employ Quinean ontology with respect to existential quantification and hence make no distinction between "there exists" and "there is."
[11] W.V.O. Quine, "On What There Is," in *From a Logical Point of View: 9 Logico-philosophical Essays* (Cambridge: Harvard University Press, 1961), 13.

to be ontologically committed to whenever we give a statement about characters.[12]

We can syllogise van Inwagen's argument as follows:

(P1) (4) expresses a true proposition.
(P2) (4*) is a correct translation of (4) into formal logic.
(P3) The rules of formal logic are truth-preserving.
(P4) (4*) commits us to the view that there are such things as characters.
(C1) Therefore, there are such things as characters.[13]

This argument seems entirely wrongheaded to me, for it seems incorrect to interpret the statement, which is given in ordinary language, as asserting a philosophical position or being ontologically committed to the existence of characters. When we translate the sentence into quantifier logic, we require the use of the existential quantifier to say "there exists an x such that x is a character in a 19th-century novel." But it seems to be incorrect that (7) is actually asserting a philosophical position that there are characters.

In ordinary language, we use expressions without necessarily committing ourselves to a certain philosophical position. Consider Jared, a prominent scientist, who is climbing a mountain with his friend. They reach the peak just as the sun is setting and Jared exclaims, "That is a beautiful sunset!" It would make no sense for his friend to respond, "You should know that sunsets don't really exist. The sun isn't actually setting, it's just how it looks because of the rotation of the Earth." Clearly it would be incorrect to take Jared as asserting that sunsets are not illusions but actually exist, even though he is using vocabulary that seems to commit him to recognizing the existence of sunsets; in fact, Jared knows that sunsets are illusions. If he were attempting to be scientifically accurate he would say, "The rotation of the Earth and its elliptical orbit around the sun creates such a beautiful illusion." The point of his original assertion is not to perfectly describe the world, but to simply express his attitude towards a certain feature of the world (and, more specifically, with how that feature

[12] van Inwagen, "Creatures of Fiction," 302.
[13] Ibid.

appears to be). Thus, in order to determine what a speaker is actually asserting when expressing a statement, we also need to understand the intentions of the speaker.

This fact becomes apparent when we examine those ordinary language statements which clearly do commit the speaker to a philosophical position or the existence of an entity. If Bill, a poorly informed individual, exclaimed, "That is a beautiful sunset!" we might be willing to claim that Bill has committed himself to the existence of sunsets. Notice that this exclamation is identical to Jared's but that, while Jared was not committed to the existence of sunsets, Bill is. The upshot is that there seems to be no property of these exclamations as such that makes them ontologically committing. The most plausible explanation, it seems, is that Jared knows his use of the word "sunset" accurately describes the world, whereas Bill falsely believes that his use of the word "sunset" accurately describes the world. These two epistemic statuses are key in determining what these two speakers intend to assert: Jared does not intend to assert that sunsets exist because he knows that they do not, whereas Bill does assert this because he falsely believes that sunsets do exist. Preliminarily, as far as ordinary language goes, ontological commitment largely seems to depend upon what speakers intend to commit themselves to and not what their statements appear to commit them to.[14]

Thus, when we express statements such as (4) in quantifier idiom, it is incorrect to assume that our language commits us to the existence of characters, just as it is incorrect to assume that Jared's exclamation has committed him to the existence of sunsets. As far as ordinary language goes, it seems fundamentally incorrect that statements commit speakers as

---

[14] The upshot of this position is that speakers can never be unintentionally committed to anything (in ordinary language). A full solution to this problem is outside the scope of the present paper, but to give a preliminary response, this fact is not as absurd as it seems at first glance. There is clearly an implicit distinction between the ontological commitment of a statement and a speaker. For a speaker to be ontologically committed, his intention must align with the ontological commitment of a statement. Thus, a speaker may accidentally utter a statement that commits himself to the existence of some entity $x$ without thereby committing himself to the existence of $x$. Take, as an analogy, an innocent man who accidentally gives a statement $P$ that commits himself to being a criminal. He has unintentionally committed himself to being a criminal by uttering $P$, but he is clearly not actually committed to this, as is evident when the man comes to realize what $P$ actually means and exclaims, "That is not what I meant; let me rephrase that."

the case between Bill and Jared demonstrates.[15] Rather statements reflect the speaker's commitments. The exclamation, "That is a beautiful sunset," is not ontologically committing in and of itself, and it only becomes ontologically committed to the existence of sunsets when speakers intend so.[16] Similarly, we cannot assume, as van Inwagen does, that (4) commits speakers to the existence of characters. As in Jared's case where we treated the word "sunset" as a circumlocution, so we may also wish to treat the word "character" in (4) as a circumlocution and give a clear paraphrase: If we examine a wide collection of 18th and 19th–century books and count the statements that give physical details about appearance, we should find more of these statements in 19th–century than in 18th–century books. Or, perhaps more accurately, we may wish to say that 19th–century books have more adjectives relating to appearance than 18th–century books. Thus, in order to be an effective argument, it seems that van Inwagen must accept a number of dubious assumptions about assertions and philosophical commitment in ordinary language.

**IV. Fictional Assertives Lack Reference**

Although I have argued that van Inwagen's argument fails to establish the existence of fictional objects, I will assume the soundness of the argument in order to assess his complete answer to the problem of reference. As such, we now turn to his analysis of the three types of statements about fictional objects: fictional assertives, literary descriptives, and meta-fictional statements.

Recall that fictional assertives are statements made by authors describing fictional objects. At first, it seems that van Inwagen's theory gives a simple analysis of how we may perform a truth-evaluation of these

---

[15] Here, I am agreeing with Searle that speaker's meaning is prior to sentence meaning. "Speech Acts and Illocutionary Logic," in Logic, Thought, and Action, ed. by Daniel Vanderveken (Dordrecht: Springer Netherlands, 2005), 117–118.

[16] This point needs clarification. Although statements in and of themselves are not ontologically committing, linguistic practices may designate that certain statements are ontologically committing. For instance, for an ancient tribe, statements made about sunsets may be ontologically committed to the existence of sunsets, and this linguistic custom defines what a speaker may do with language. If a speaker, in this community, does not wish to commit himself to the existence of sunsets, then he must make this apparent. If he states, "That is a beautiful sunset," his statement is committed, by custom in his community, to the existence of sunsets even though he intends to not be committed.

sorts of statements; for when Doyle states that Sherlock Holmes was a detective or lived on 221B Baker Street, he is literally referring to the abstract entity Sherlock Holmes and saying something true about him. The obvious problem with this, as van Inwagen himself admits, is that abstract entities cannot have concrete properties like being spatially located.[17] If Sherlock Holmes is an abstract entity, then Doyle says something false when he states that Sherlock Holmes is a detective, but this clearly cannot be correct.

Van Inwagen's solution is simply to reject that fictional assertives are about creatures of fiction at all. When authors describe their characters in stories, they are neither actually making a claim nor writing about anything. When Doyle gives any fictional assertive about Sherlock Holmes (that he is a detective for instance), this does not represent an attempt at saying anything about Sherlock Holmes, since Doyle was neither referring to nor asserting anything. For this claim, van Inwagen relies upon an argument given by J. O. Urmson.[18]

According to Urmson, authors who write fictional assertives make no attempt at asserting a proposition about the world, and so it makes no sense to evaluate these statements as true or false. To give Urmson's analogy, suppose Carl and Nigel decide that they are going simulate the 1994 game between Short and Kasparov, with Carl playing as Short, and Nigel playing as Kasparov. They proceed to play the game using exactly the same set of moves that their respective 1994 counterparts used. In this case, it makes no sense to ask the question of Carl and Nigel, "Who won?" Neither of them won because neither of them were actually playing a real game of chess. Although they abided by all of the rules, and although it may have looked perfectly 0like a real game of chess to any bystander who happened to be present, no real game of chess was actually played. As Urmson says, "In the case of fiction 'Is it true?' will be inappropriate for the same reason as 'Who won?' is inappropriate to the mock-chess."[19]

Nevertheless, because of how similar the mock-chess game is to the 1994 game, it is quite natural to describe the mock game in the

---

[17] van Inwagen, "Creatures of Fiction," 305.

[18] For a similar argument, see Alvin Plantinga, *The Nature of Necessity* (Oxford: Clarendon Press, 1988), 149–159.

[19] J. O. Urmson, "Fiction," *American Philosophical Quarterly* 13, no. 2 (1976): 156.

language of real chess. If Nigel (playing as Kasparov) moves his queen pawn from D2 to D3, it is sensible enough to borrow this vocabulary as opposed to some odd "pretend" or "mock-chess" language.[20]

Similarly, just as statements made about mock-chess are not really statements made about chess, fictional assertives made by authors about creatures of fiction are not really about creatures of fiction. In fact they are not even assertives at all; they are pretending to be assertives merely for the pragmatic value of avoiding the use of some odd fictional language. According to van Inwagen, fictional assertives do not even "represent an *attempt* at reference or description."[21] If correct, this provides van Inwagen with a very simple analysis of fictional assertives. When read literally, any fictional assertive that Doyle has given is false, since abstract entities (like the character Sherlock Holmes) cannot have concrete properties (like living on 221B Baker Street). But as it turns out, when Doyle gives the fictional assertive, "Sherlock Holmes lives on 221B Baker Street," this statement does not refer to the abstract entity "Sherlock Holmes," since it is not really a statement about anything. Though fictional assertives appear identical to real assertives, they differ in their lack of reference and truth-evaluability.[22]

I think that the problem with this claim about the non-referential nature of fictional descriptives is that it is based upon an insufficient theory of reference. Both van Inwagen and Urmson appear to be advocating for the following definition of an assertive:

A sentence *A* is an assertive iff (i) *A* has a descriptive propositional content *p*; (ii) *p* does not contain any empty terms (i.e., every term in *p* has a reference); and (iii) a speaker *S* who asserts *A* is committed to the truth of *p*.

The problem arises when we consider that assertives presuppose that every term in *p* already has a designated reference. Assertives are not the type of speech act which can fix reference since, by the above definition, they are merely descriptive, whereas the designation of a reference involves an intentional action. When I look out my window and

---

[20] Ibid.
[21] van Inwagen, "Creatures of Fiction," 301.
[22] Urmson "Fiction," 155.

assert, "The weather is stormy outside," I am simply describing the way the world is or appears to me. This is quite different from what is going on when the parents of a newborn declare for the first time, "Our baby is called 'Mary'" or when a ship owner announces, "I dub this ship 'Challenger.'" These sorts of statements do not describe the world as it is, for newborns and ships do not come with pre-assigned names. Rather, these statements attempt to linguistically change the world; the parents and the owner have not described reality, for the names "Mary" and "Challenger" do not have references until the parents and the ship owner respectively designate them. Upon assigning these names, they have created new linguistic facts about the world.[23] Following Searle, I will take to calling these sorts of statements "declaratives."[24]

Often times, statements which appear to be assertives are really declaratives. For instance, imagine that in the distant future, a group of astronauts leave their homes on Earth and set sail to find a new planet to colonize. Far outside our solar system, they happen upon a habitable planet and land on one of its many islands. After deciding to permanently stay on that island, they state, "This island is a country called 'Atlur.'" Of course, this statement appears to be an assertive which describes a certain state of affairs, for it has the same form as other assertives like, "This is a rock," or, "This chair is made of plastic." Countries, however, are social phenomena, and so in order to classify that specific island as a country, an initial declaration is required. After the declaration has been made and the reference fixed, these statements become assertives. Prior to its establishment, "Atlur" was an empty name since it had no reference or content. After the declaration, "This country is called 'Atlur,'" the reference of the name was fixed to a specific area of land. After this baptism, when a person now states, "This country is called 'Atlur,'" he is performing an assertion, not a declaration, since he is describing the world

---

[23] This is based off of Kripke's theory of naming. See Saul Kripke, "Naming and Necessity," in *Semantics of Natural Language*, ed. Donald Davidson and Gilbert Harman (Boston: D. Reidel Publishing Company, 1973), 290–293

[24] John R. Searle, *Expression and Meaning: Studies in the Theory of Speech Acts* (Cambridge: Cambridge University Press, 1979), 16–19.

and not fixing a reference.[25]

Although fictional assertives appear to be actual assertives, they are in fact declaratives. When Doyle writes statements about Sherlock, he is not describing the world but fixing the reference of the name "Sherlock Holmes." Of course, in order to fix reference, there must be some sort of subject of this baptism. In our case of the country, the settlers fixed the reference to a particular island, but Doyle clearly cannot fix the reference of "Sherlock Holmes" to a concrete object. As we have seen, van Inwagen takes it that creatures of fiction are abstract objects and that fictional assertives do not refer to them or anything else. In contrast, I think it makes more sense to think of creatures of fiction as mental objects, and that these objects are the subject of reference.

Firstly, unlike abstract objects, fictional objects clearly have temporal and ontological beginnings. As the author, Doyle quite literally brought Sherlock Holmes into being at a certain time, a fact that van Inwagen clearly concedes.[26]

Secondly, abstract and fictional objects have a quite different ontology altogether. For the latter are existentially dependent upon other entities, whereas the former are not. Roughly, an entity $x$ is existentially dependent upon another entity $y$ if $x$ could not exist unless $y$ exists. For example, concrete objects such as mountains and trees are existentially independent, whereas countries, currency, marriage, and Presidents (which we might call social objects) are existentially dependent upon persons. The property of being currency is not a property of the physical world and so has no independent existence outside of the persons who conceive of it. The piece of paper and ink only becomes currency when persons define it as such. Likewise, as traditionally conceived, abstract objects are existentially independent.[27] In contrast, creatures of fiction are clearly existentially dependent, for if no person ever conceived of Sherlock Holmes, he would never have existed.

---

[25] Though Kripke does not make this distinction explicitly, it is certainly implicit in his theory of naming.

[26] van Inwagen, "Creatures of Fiction," 306.

[27] Bob Hale, "Realism and Antirealism about Abstract Entities," in *A Companion to Metaphysics*, ed. by Jaegwon Kim, Ernest Sosa, and Gary S. Rosenkrantz, (Oxford: Blackwell Publishing, 2009), 66.

We can now answer the question of what Doyle fixes the reference of "Sherlock Holmes" to when he makes a fictional assertive. These statements are declaratives (although they appear to be assertives) which fix the reference to mental objects and have the following form:

(6) If something is an idea of a man who is a detective who lives on 221B Baker Street and has further properties $P_1$ through $P_n$, then it is a mental object of Sherlock Holmes Where $P_1$ through $P_n$ are properties designated by Doyle.

Once again, Doyle's first use of a statement with the form of (6) is a declarative and not an assertive, for it is fixing the reference of "Sherlock Holmes" to a type of mental object.[28] Now we can see how fictional assertives can be about their objects of reference, for they do not presuppose that the reference is already fixed, but they instead establish the reference themselves. They are really "fictional declaratives." Consequentially, a fictional declarative is truth-evaluable, though of course, the statements are trivially true. Returning to the previous example, when the astronauts first give the declarative that the island they are standing upon is a country called "Atlur," they could have equally used any other name and still produced a true statement.

Obviously, there is nothing special about the specific name that they use. The only relevant fact is that they do in fact designate a name for the island, and whatever name they pick, it is automatically true (at the moment of naming) that the island is called that name.[29] Since a declaration creates a fact about the linguistic representation of the world, it is true the moment it is uttered, just as any fact of the world is true the

---

[28] Note that these statements are only declaratives the first time Doyle makes them; afterwards, they become assertives. For instance, if Doyle says, for the first time, "Sherlock Holmes is a detective," this statement is a declarative. If, the next day, Doyle says the same statement to a friend, he has given an assertive whose truth value is dependent upon the initial declarative.

[29] Of course, this in no way implies that assertives will be permanently true. If the astronauts decide at $t_1$ that the island is called "Atlur" and, at $t_2$, agree to change the name to "Antonia", it will be true that, at $t_1$, the island was called "Atlur," and, at $t_2$, the name of the island was changed to "Antonia."

moment it obtains.[30] A similar story may be told about fictional assertives as well: for any name *x* that Doyle picks for his character, it is true, at the moment of naming, that the character is called *x*.

While I am convinced that fictional objects have their reference fixed by declaratives, I am uncertain as to what sort of thing fictional objects really are. Although I have stated that they are some kind of mental object, this is merely an inference to the best explanation, as it seems apparent that they cannot be abstract or concrete objects. Though I have argued against fictional objects being a kind of abstract object, I have not explicitly said anything about fictional objects being concrete. However, since fictional objects are existentially dependent entities, they cannot be concrete objects, which are existentially independent.[31]

**V. Ascription and Predication**

The final part of van Inwagen's theory of fictional objects addresses his analysis of literary descriptives and meta-fictional statements. Note that van Inwagen's answer regarding fictional assertives does not address this problem, for a statement is a fictional assertive only when it is given by the author. When a speaker other than Doyle utters that Sherlock Holmes is a detective, this is not a fictional assertive, but rather a literary descriptive, a statement made by speakers other than the author and about creatures of fiction that have already been written about. Rather, according to van Inwagen, when analyzing literary descriptives and meta-fictional statements, the key is distinguishing between types of properties and types of relations fictional objects can have with properties.[32]

In addition to all of the concrete properties that seem to be predicated to fictional objects, there are all sorts of literary properties that fictional objects seem to have as well. These properties include being a character in a novel, being created by an author, being the main villain, and so forth. The solution to this puzzle, according to van Inwagen, is that

---

[30] This is often referred to as a "double direction of fit." See. Daniel Vanderveken. *Meaning and Speech Acts Volume 1: Principles of Language Use.* (Cambridge: Cambridge University Press, 1990), 106.

[31] Jay E Bachrach, "Fictional Objects in Literature and Mental Representations," *British Journal of Aesthetics* 31, no. 2 (1976): 134–139.

[32] van Inwagen, "Creatures of Fiction," 305.

fictional objects have certain literary properties predicated to them, while concrete properties can only be ascribed to them.[33] Van Inwagen defines this relation of ascription "*x* is ascribed to *y* in *z*," symbolised by the three-term predicate *A*(*x*, *y*, *z*), where *x* is a property, *y* is a creature of fiction, and *z* is a work of fiction or a place in that work (for example, Chapter 2).[34] So, when we say, "Sherlock Holmes is a detective," what we really mean is "Sherlock Holmes is ascribed the property of being a detective," and that is symbolized as *A*(detective, Sherlock Holmes, Chapter 1 of *A Study in Scarlet*).

Although van Inwagen is willing to treat the relation of ascription as primitive,[35] he claims that Cartesian dualism provides a useful analogy for this relation. Someone may argue against the dualist that his position must be false, for the dualist identifies the person as an immaterial soul, yet persons obviously have concrete properties (e.g., being a certain height) while immaterial souls do not. The dualist will certainly agree that a person does not have these concrete properties. The person, instead, has mental properties such as the ability to think or to feel emotions. But souls do have the property of animating a body, and thus bear an intimate relationship to concrete properties. So while Christopher (who is identical with an immortal soul) does not *have* the property of being 175 pounds, he does have the property of animating a body which is 175 pounds.[36]

In the same way, van Inwagen argues that fictional objects have literary properties but not concrete properties. However, just as Christopher does not *have* but *bears an intimate relationship with* certain concrete properties, so too do creatures of fiction not *have* but *bear an intimate relationship with* certain concrete properties.[37]

Van Inwagen states,

> "And just as, on the Cartesian view, we may say 'Jones is six foot tall' and be talking about an immaterial substance

---

[33] Ibid., 305–306.
[34] Ibid.
[35] Ibid., 306.
[36] Ibid.
[37] Of course, the notion of "ascription" is not perfect a word, and in fact, may be quite misleading. Van Inwagen admits this, but for lack of a better term, chooses to simply accept the consequences, but not without some clarification (see van Inwagen, "Creatures of Fiction," 305–306).

> without thereby predicating being six foot tall of that immaterial substance, so, on the present view, we may say 'Mrs. Gamp is fond of gin' and be talking about a theoretical entity of criticism without thereby predicating fondness for gin of that theoretical entity of criticism."[38]

If correct, this is quite an ingenious solution to the problem. However, the reason why this solution works for Cartesian dualism and not for fictional objects is that the dualist can say that an immaterial soul bears a relationship to concrete properties because that soul is said to *animate* the body which *has* those properties. But on van Inwagen's account, what feature of a creature of fiction allows him to similarly say that it too bears an intimate relationship to certain physical properties? In order to be ascribed a property, there must be some connection between the creature of fiction and the thing to which the property in question is predicated, just as in the case of Cartesian dualism. However, fictional objects, on van Inwagen's conception, are abstract entities, so in what way could they have such a connection?

One solution may be to appeal to possible world semantics. Hence, we may take the sentence:

(2) Sherlock Holmes is a detective who lives on 221B Baker Street

and analyze it as stating,

> (2*) There is a possible world *W1* in which Sherlock Holmes is a detective who lives on 221B Baker Street.[39]

Thus, the reason why Sherlock Holmes can be *ascribed* the properties of being a detective and living on 221B Baker Street is because there exists a possible world in which he actually does exist and is *predicated* the aforementioned properties. Thus, the abstract Sherlock Holmes bears an intimate relationship to the Sherlock Holmes in *W1* who is predicated of these properties.

For a moment, let us assume with van Inwagen that fictional objects are abstract. According to possible worlds semantics, when we say

---

[38] Ibid., 305.

[39] For a similar solution, see Graham Priest, *Towards Non-Being. The Logic and Metaphysics of Intentionality* (Oxford: Clarendon Press, 2005), 116–125.

that a proposition is possible, we mean that the proposition is true in at least one possible world. The question is how we can analyze fictional objects using possible world semantics without committing ourselves to modal realism. Follow the traditional Kripkean account, possible worlds are simply abstract logical devices.[40] However, the whole purpose of analyzing (2) in terms of possible world semantics was to explain how something that is abstract could have properties ascribed to it. But our analysis of (2) as (2*) is certainly no solution, for we end up having to explain our theory of fictional objects by appealing to yet another abstract object. If this relation of ascription fails to apply to fictional objects, then on van Inwagen's account, we lack any ground for saying that literary descriptives (like, "Sherlock Holmes is a detective") and meta-fictional statements (like, "Sherlock Holmes is more famous than any detective today") can be truth-evaluable.

**VI. Concluding Remarks**

In sum, we have seen how the three parts of van Inwagen's model fail. First, his quantificational argument for the existence of creatures of fiction posits an infeasible interpretation of statements of ordinary language. Second, his idea that fictional descriptive sentences are neither an attempt to refer to anything nor truth-evaluable is due to a false classification of assertives (which presuppose that reference is fixed). In fact, these sorts of sentences are declaratives and so serve themselves to fix the reference. Lastly, van Inwagen's idea that fictional objects can only have literary properties predicated to them and concrete properties ascribed to them (an idea which he uses to give a truth-analysis of literary descriptives and meta-fictional statements) fails, since fictional objects are not the types of entities which can have properties ascribed to them. From these criticisms, I have argued that we can begin to build a rough model about fictional objects, where authors give declaratives to fix the reference of names like "Sherlock Holmes" to mental objects. Although I have my reservations about whether fictional objects could be mental objects, it seems that this is currently the best explanation, though there is certainly more metaphysical groundwork that is required here.

---

[40] Kripke, "Naming and Necessity," 267.

**Works Cited**

Bachrach, Jay E. "Fictional Objects in Literature and Mental Representations." *British Journal of Aesthetics* 31, no. 2 (1976): 134–139.

Doyle, Arthur Conan. *A Study in Scarlet*. New York: Modern Classics Library, 2003.

Hale, Bob. "Realism and Antirealism about Abstract Entities." In *A Companion to Metaphysics*, edited by Jaegwon Kim et al., 65-73. Oxford: Blackwell Publishing, 2009.

Inwagen, Peter Van. "Creatures of Fiction." *Philosophical Quarterly* 14, no. 4 (1977): 229–308.

Kripke, Saul. "Naming and Necessity." In *Semantics of Natural Language*, edited by Donald Davidson and Gilbert Harman, 253–355. Boston: D. Reidel Publishing Company, 1973.

Plantinga, Alvin. *The Nature of Necessity*. Oxford: Clarendon Press, 1988.

Priest, Graham. *Towards Non-Being. The Logic and Metaphysics of Intentionality*. Oxford: Clarendon Press, 2005

Quine, W.V.O. "On What There Is." In *From a Logical Point of View: 9 Logico-philosophical Essays*, 1-19. Cambridge: Harvard University Press, 1961.

Searle, John R. *Expression and Meaning: Studies in the Theory of Speech Acts*. Cambridge: Cambridge University Press, 1979.

—— and Vanderveken, Daniel. "Speech Acts and Illocutionary Logic." In *Logic, Thought, and Action*, edited by Daniel Vanderveken, 109-133. Dordrecht: Springer Netherlands, 2005

Urmson, J. O. "Fiction." *American Philosophical Quarterly* 13, no. 2 (1976): 153–157.

Vanderveken, Daniel. *Meaning and Speech Acts Volume 1: Principles of Language Use.* Cambridge: Cambridge University Press, 1990.

# Interpersonal Preference Comparison

Enhua Hu

## Introduction

This paper discusses a way of formalizing our intuitions concerning interpersonal preference comparisons for pairs of agents. The first section discusses previous literature in the field. The second section presents the theory informally while justifying its philosophical foundations. The third section shows that the theory can be formalized as an ordering which is complete and transitive given some restrictions. The fourth section shows that the ordering can give rise to a Pareto-optimal social choice function with advantages over traditional ones.

A few concepts need clarification. The paper works within the standard framework of microeconomics, which studies individual behavior and preference. States are represented as vectors $(x_1,x_2,x_3...)$ of $R^n$, in which each number of the vectors indicates the quantity of a particular good. Each agent has a preference ordering (denoted by $\precsim$) over the states. Rational choice theory further assumes that the preference ordering is complete ($x \precsim y$ or $y \precsim x$) and transitive (if $x \precsim y$ and $y \precsim z$ then $x \precsim z$). Lastly, most preference ordering can be represented by utility functions; this is particularly interesting if the utility functions are continuous.

## I. Survey of Literature

Interpersonal comparison (IC) statements have the form, "A prefers x more than B prefers y". Many economists believe that an IC statement cannot have a truth value because it uses an ordinal concept of utility, which ranks the preferences but does not capture the intensity of the preferences. Thus, if there is no way of comparing intensity for even single agents, it seems nonsensical to compare intensity between agents.[1]

[1] D. M. Hausman, "The Impossibility of Interpersonal Utility Comparisons," *Mind.* 104:473–490, 1995.

Further, the utility of goods is subjective; there is no objective measurement we can use between agents.[2] More recently, Rossi has proposed an epistemological argument against the possibility of ICs.[3]

Despite this view, I argue that the possibility of IC is real. In addition to the theoretical benefits,[4] we have a common intuition that some ICs are true; we know that a starving child gets more utility from bread than does an adult who is full. If the possibility is denied, then statements like the above have no truth value, which is highly counter-intuitive.

The section will survey two main theories for IC and point out their issues. They are:

- Extended sympathy, advocated by Harsanyi[5] and others.[2]
- Zero-one rule, introduced by Isbell;[6] Hausman[1] is a proponent.

*Extended Sympathy/Empathetic Preference*

Arrow[5] and Sen[6] call this theory 'extended sympathy', Harsanyi calls it 'empathetic preference', and it is by far the most studied theory of IC. The assertion of empathetic preference is that individuals can empathize with others. Theorists agree that there are generally two processes that must be accomplished for an individual to have a judgment of extended sympathy. First, the individual must undergo an objective shift where she considers herself in the material condition of another agent. Second, the individual undergoes a subjective shift where she replaces her preferences with those of the other agent. These two shifts allow an agent to capture the preference of another via empathetic understanding.

In the following, I will present Harsanyi's version of the concept. Harsanyi's model has two axioms:

---

[2] L.Robbins, "Inter-personal Comparisons of Utility: A Comment," *Economic Journal,* 48:635–641, 1938.

[3] M. Rossi, "Interpersonal Comparison of Utility. The Epistemological Problem," (PhD Thesis, London School of Economics and Political Science, 2009).

[4] K. Arrow, "Extended sympathy and the possibility of social choice," *Philosophia,* 7:233–237, 1978.

[5] J. Harsanyi, "Cardinal welfare, individualistic ethics, and the interpersonal comparison of utility," *Journal of Political Economy,* 63:309–321, 1955.

[6] J. R. Isbell, "Absolute Games," in A. W. Tucker and R. D. Luce (eds) *Contribution to the Theory of Games, Vol. IV*, Princeton University Press, 357–96.

1. Utility functions, empathetic and personal, satisfy all Neumann-von Morgenstern postulates, such as completeness of preferences.
2. Agents are able to empathize fully with one another and they are able to capture the exact same preference relation. In technical terms, an agent i's empathetic function for agent j is a strictly increasing affine transformation of agent j's own preference function.

Harsanyi allows for an agent to maximize the sum of all her empathetic functions by evaluating how much the agent values the utility of one over another agent, a form of IC. Each empathetic function is subjective; therefore, this agent cannot be an impartial social planner as she is. Harsanyi deals with this by putting this agent in a situation of uncertainty such that she must make her decision as if she has equal probability to be any other agent. The idea is similar to Rawls'[11] veil of ignorance, only that Harsanyi makes the agent maximize expected utility rather than follow the maximin principle.

Each empathetic function is subjective and needs to be so since it is one's own IC. Therefore, different agents will likely make different choices when put in this situation. Harsanyi is aware of this; his argument is that agents with enough information will have the same empathetic functions. Second, the idealizations required by Harsanyi are very demanding. In particular, the ability to empathize perfectly and to be an impartial observer may only be conceptually possible.

*Zero-One Rule*

The zero-one rule was first introduced by Isbell[7] as a criterion of fairness. Unlike empathetic preference, this rule employs utility as cardinal and bounded. The idea is to normalize an agent's utility function such that the utility of the most preferred state is 1 and the utility of the least preferred state is 0.

The theory says that agent i is better off in state x than agent j in state y, iff the following condition holds:

$$\frac{MaxU_i - U_i(x)}{MaxU_i - MinU_i} > \frac{MaxU_j - U_j(y)}{MaxU_j - MinU_j}$$

The above can be interpreted as one agent being "closer" to her best state than another. This interpretation only makes sense with cardinal utility because it requires utility to represent intensity. Furthermore it implies that there are most/least preferred states. While these requirements are demanding, the main problem is the hidden assumption that agents have equal capacity for preference. The theory implies that agent i and j are equally well off in their worst and best states.[1] Sen argues that a social welfare resulting from this would emphasize society's resources with lower satisfaction requirements. Given limited goods, the social planner wanting to maximize utility will have to allocate to those who are satisfied with less.

The response from the zero-one rule proponent is to argue that preference satisfaction is what the theory compares, rather than well-being. They argue that well-being is not being compared; rather, the theory compares preference satisfaction. Hausman states:

> "No sense has been given to comparing Jill's non-comparative well-being to Ira's non-comparative well-being. In the case of cardinal and bounded utilities, the conclusion ought to be that a view of well-being as preference satisfaction leaves interpersonal comparisons undefined and mysterious."[3]

The argument is that the normalized utility function only describes how well preferences are satisfied and that preference satisfaction is not well-being. The theory evaluates statements of the following form: "Individual i's preferences in state x are better satisfied than individual j's preferences in state y."

If one takes the preference satisfaction approach, then this theory seems more reasonable, albeit less powerful because it has little implication for social welfare. If all we can compare is the extent of preference satisfaction and not the strength of satisfaction itself, then we cannot maximize social welfare. The zero-one rule still leaves a hole to be filled, namely that there is no way of comparing welfare between individuals such that we can assign truth value to statements of the form, "individual i is better off in state x than individual j".

## II. Philosophical Issues and Informal Layout

This section lays out informally my theory of IC statements. This exposition captures some of the justifications and intuitions for the theory; I address the more technical and formal concerns in the next section.

The two main problems that face ICs are the lack of objective basis for comparison and the ordinality of utilities. The extended sympathy method is flawed because the objective basis provided is weak. The zero-one rule is flawed because it requires the cardinality[7] of utilities, which goes against standard economics.

My theory deals with the second problem by comparing only preference orders, so the theory always uses ordinal utility. The first problem is more intricate; it might very well be that no objective basis can be found for IC, so my theory proposes intersubjective agreements between agents as the basis for IC. More precisely, statements of the form, "Agent A prefers x more than agent B prefers y" are true if "A thinks she prefers x more than B prefers y" and "B thinks she prefers y less than A prefers x" are true. The claim is that agreements between subjective judgments can entail the truth-bearing quality of IC statements.

### *What is preference?*

IC statements have the form "A prefers x more than B prefers y"; they have two objects and one binary connective. The objects are the preferences the agents have for a good. In standard economics, preference is defined as a relative ranking over the set of goods. One approach would be to define preference for a particular good as the ranking of the good, i.e., the fifth most preferred good. However, this approach fails because many rankings have infinite goods and no most preferred good.[8]

I propose that an agent's preference for a good is described as the set of all goods to which the agent is indifferent, their 'indifference set' for

---

[7] In economics, utility is conceived as either cardinal or ordinal. If cardinal, then the number represents intensity or degree of preference, namely if an apple gives five utility, a pear ten, and a banana three, then we can say that the pear is preferred to the apple *more* than the apple to the banana. If ordinal, then the numbers simply denote the ranking or order of preference. Therefore, all that can be said is that the pear is preferred to the apple which is preferred to the banana.

[8] For instance, a ranking for quantities of a continuous good like water.

that good. When we ask someone how much they value a good, their preference, we ask them to provide an equivalent good. However, we cannot describe an agent's preference for a good, x, with just one equivalent good, y. It may be the case that another agent's preference with respect to the two goods (x,y) are the same but differ for a third good, z. If preference is described with respect to only one equivalent good, then the two agents would have the same preference for x with respect to y but not with respect to z. However, preference for a good cannot be subject to other goods but only agents, so we must look at the whole indifference set rather than any particular member when defining preference.

*Subjective Judgments*

My proposed method uses intersubjective agreements to analyze statements of IC. Subjective judgments have the form, "A considers her preference for x to be stronger than B's preference for y." Once again, this is a binary connective with two objects. The two objects are first, A's indifference set for x, and second, B's indifference set for y.

Since this is A's subjective judgment, this binary connective should be A's preference ordering. A's preference ordering allows her to compare individual goods, but gives her no way of comparing sets of goods. I propose that preference over sets of goods be defined as a range, this range defined by the most and least preferred good in that set.[9] Let me clarify this with the following example.

Imagine an agent who is asked to compare her preference for a good, x, and another agent's preference for a good, y. She will look at the members of the other agent's indifference set for y. If it is the case that she considers all those goods superior to x, then she concludes that the other agent values y more than she values x. If our agent considers all those goods inferior to x, then she concludes that the other agent values y less than she values x.

Another case is one in which there are goods in the other agent's indifference set that our agent considers more and less valuable than x. This case has our agent believing her preference to be vaguely similar to

[9] This is slightly technical; I elaborate this further in section III.

her counterpart's. This type of scenario, while unappealing at first, is intuitive and arises often. For instance, two children who like certain chocolates are asked who likes them more.

One might ask, "Why would an agent ever think that her preference is weaker than that of another?" One may think that the agent is not empathetic, or that she could report falsely. A lack of empathy is not a concern because subjective judgments do not require empathy. For false reporting, while most social choice theorists assume to know the true preference of agents, it may not be the case. This is also known as the "Preference Revelation Problem," and there are mechanisms designed so that the agent is forced to report her true preference as it is always her optimal action.[10] This paper will not consider the problem of preference revelation, as it is beyond its scope; instead, I will take the orthodox view that agent preferences are available.

*Intersubjective Agreement*

Now that I have presented subjective judgments, I will argue that the truth of an IC statement depends on the subjective judgments of the agents whom the IC statement concerns. In particular, if two agents' subjective judgments agree with each other, then the IC statement is true.

This claim is hard to verify because there are no standards for the truth of IC statements. In the following, I propose two views: Either IC statements are equivalent to intersubjective agreements, or the weaker alternative, the truth of intersubjective agreements implies the truth of IC statements.

Recall the problem of objective basis. I will show that in the absence of an objective basis, intersubjective agreements are equivalent to ICs. If there is an objective basis, then intersubjective agreements merely imply IC statements.

Suppose there is no objective basis for IC statements. Then, for the sake of social choice, we must still decide how to best assign truth values for those statements. If it is the case that statements of IC are subjective, then whose subjectivity matters? Clearly, the agents whom the

[10] The most famous one being the Vickrey-Clark-Groves Auction.

IC statement concerns have priority. With the absence of an objective ground, the best judgment is one that both agents agree to. If the two agents agree that one has a stronger preference, then an outsider's judgment should not matter. Of course, the two agents may not always agree since not all IC judgments need to be true.

Suppose that there is an objective basis for IC statements, such that there are objectively true and false IC statements. Then individual agents, given enough information, can arrive at the right conclusion regarding ICs. Some statements of IC have seemingly immediate truth interpretations, for instance, "A starving person's preference for food is stronger than a satiated person's preference for food." Other IC statements require more information. Clearly, the agents whom the IC statements concern have the most information, as they know their preference orderings best. Therefore, if the two agents' judgments agree, it is the best approximation of the objective truth. This is similar to what Harsanyi claims, except my method does not require an omniscient social planner.

If one accepts my arguments above, it follows that intersubjectivity allows the analysis of IC statements. One might still ask whether there is an objective basis for ICs. We are now in a position to give some insight into that question. Since, in the absence of an objective basis, IC statements are equivalent to intersubjective agreements, if there are IC statements which are true but do not obtain intersubjective agreements, there must be an objective basis. Unfortunately, we cannot answer this question any further. Since we do not have a formal definition of truth for IC independent of intersubjectivity, we must rely on intuition. Therefore, the set of IC statements which are true will be the intuitively true and obvious ones which will likely always obtain intersubjective agreements. While there may be non-obvious but true IC statements which do not obtain intersubjective agreement, my method cannot analyze them.

## III. Formal Language and Interpersonal Preference Order

In this next section I present a logic capable of expressing preference orderings of pairs of agents and I show that we can build an ordering of interpersonal preference from it. Further, this ordering is complete and transitive if the single agent preference satisfies some basic properties.

*Syntax*

The syntax of two agent preference logic is the following:

- The usual logical symbols of predicate logic.
- A set of goods/states: S = $\{x, y, z\ldots\}$
- A set of binary relations over S, P = $\{\precsim_1, \sim_1, \prec_1, \precsim_2, \sim_2, \prec_2\}$
- A set of unary relations over S, Q = $\{Q_1^1, Q_1^2, Q_1^3, \ldots, Q_n^n\}$

A well-formed formula (wff) is defined the same way as in predicate logic.

*Semantics*

The set of states consists of bundles of goods in $R^n$, so each unary relation $Q_i^j$ denotes that a good has j quantity of the $i^{th}$ component.

Example: $Q_2^4x$, denotes that $x$ has 4 of the 2nd component.

The binary relations capture the preference ordering of two agents, 1 and 2, over the set of states, their meanings are strict/weak preference and indifference.

Example:
$x \sim_1 y$, denotes that agent 1 is indifferent between $x$ and $y$
$\forall x(\forall n(Q_n^0 x \rightarrow \forall y(x \neq y \rightarrow x \prec_1 y)))$, denotes that if $x$ is an empty state (zero in all components), then any good $y$ is strictly preferred over $x$ by agent 1.

In standard economics, it is assumed that all three relations are transitive. However, $\prec$ and $\sim$ are not complete. Only $\precsim$ is complete, transitive, reflexive and symmetric. Lastly, $\sim$ is an equivalence relation while $\prec$ is a strict total ordering. Their relationship is as follows:

$$(x \sim y) \leftrightarrow (x \precsim y \wedge y \precsim x)$$
$$(x \prec y) \leftrightarrow (x \precsim y \wedge \neg (y \precsim x))$$

*Some definitions*

I now move on to the construction of the interpersonal ordering. First, I define the concept of indifference sets for the evaluation of agent preference over a good. Second, I define interval sets, which captures the concept of subjective judgments.

Def.1 Indifference Set:
Let $x$ and $i$ be respectively a bundle of goods and an agent. Denote $[x]_i$ to be the indifference set of $x$ by agent $i$ such that:

$$[x]_i = \{ y \mid x \sim_i y \wedge x \neq y \}$$

The indifference set of a bundle of goods, $x$, is the set of all goods the agent considers equivalent to $x$, excluding $x$ itself.

The indifference set here is different from standard definitions because I exclude the original good. It would be circular if I defined an agent's preference for a good via use of that good. Further, excluding the original good will allow for more equivalence classes in our interpersonal ordering, making it stricter. However, in the context of continuous utility functions, this assumption is not necessary.

Def.2 Interval Sets
Interval sets can be understood as ways for one agent to evaluate the preferences of another agent.

We say that $[y,z]_j$ is the interval set for agent j on indifference set $[x]_i$ iff

$$\text{For } \forall x \in [x]_i, \forall y \in [y]_j, \forall z \in [z]_j, y \precsim_j x \precsim_j z, \text{ for agent j}$$

Where the indifference sets $[y]_j$ and $[z]_j$ contains respectively the least and most preferred goods in $[x]_i$ by agent j. The interval set is interpreted as the subjective evaluation of the preference of one agent by another.

Theorem 1

For any non-empty indifference set, there exists one interval equivalence for each other agent.

The proof is shown in the construction of interval sets. Take the most and least preferred outcome in the indifference set by the other agent and assign them as boundaries of the interval. We can do that because the set is not limited by constraints but rather denote possible states.

We have now all the tools needed to define subjective judgments of the form "A considers his preference for x to be stronger than B's for y".

Def 3. Subjective Judgments of Preference

Agent $j$ considers his preference for good $w$ stronger than that of agent $i$ for good $x$, denoted by $[x]_i \prec_j [w]_j$, iff:

$$\forall x \in [y,z]_j \text{ , we have } w \prec_j x,$$

where $[y,z]_j$ is the interval set of $[x]_i$ for agent j

In other words, she prefers $w$ to all goods in her interval equivalence bundle for $[x]_i$. An illustration helps one grasp the concept and will be useful as we develop it further.

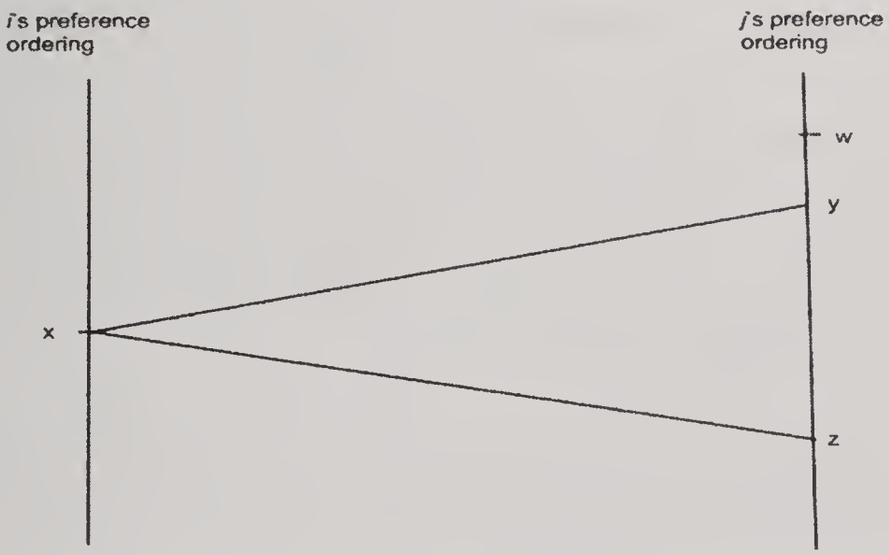



Fig.01: $[x]_i$'s interval equivalence $[y,z]_j$

We can analyze intersubjectivity now that we have defined subjective judgments. There are generally three cases of interval comparison available. The respective intervals can be exclusive from the goods and allow for either intersubjective agreement or disagreement, or the intervals are inclusive of the goods. We can easily interpret the first two cases of exclusivity. If there is intersubjective agreement then we have a strict stronger/weaker preference relation. If there is intersubjective disagreement then the social planner's job is simple since one agent wants what the other one does not want. Graphically, for the two cases we have:

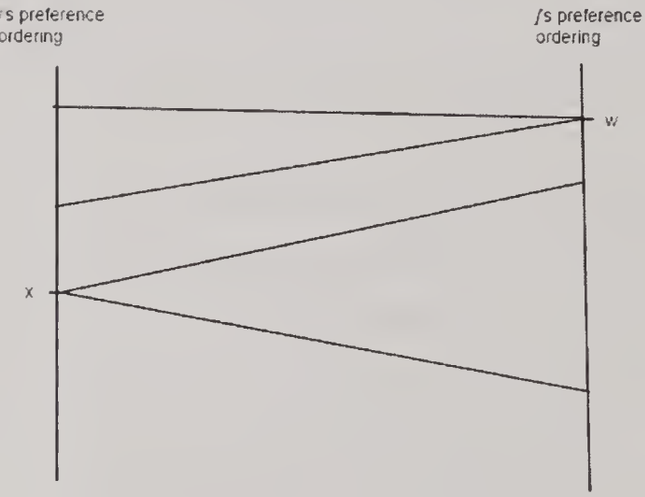



Fig.02 Intersubjective Agreement

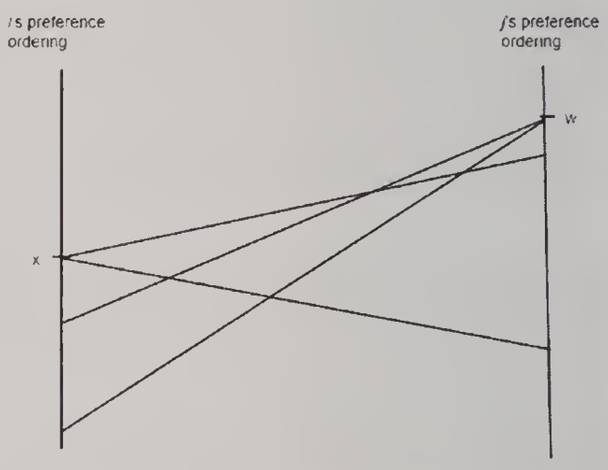



Fig.03 Intersubjective Disagreement

I interpret the case of intersubjective agreement (Fig.02) as "Agent *i* prefers *x* less than agent *j* prefers *w*." For the case of intersubjective disagreement, no interpersonal comparison can be made. Therefore, the interesting cases for the framework exclude those in Fig.03. Fortunately, two mild restrictions on individual agent preference orders exclude the case of intersubjective disagreement.

Def. 4 Strong Monotonicity[11]

Strong monotonicity is a standard axiom in economics; it is defined as follows:

Let $x$ and $y$ be bundles of n goods represented as $(x_1,...,x_n)$ and $(y_1,...,y_n)$. We say that an agent's preference is *strongly monotonic* iff the following holds:

> If $\exists x_i, y_i$ such that $x_i > y_i$ and $\forall x_i, y_i$ are such that $x_i \geq y_i$ then $x$ is preferred to $y$.

The strong monotonicity property implies that our bundle of goods has, well, goods. More formally, it means that each unit of a good has positive value for the agent.

Def. 5 Income in Bundle

For any bundle of goods, $x$, $\exists a \in \Re$ such that $y=(a,0,0,.....,0)$ and $y \sim_i x$ for all agents $i$.

This just means that our agent can equivocate any bundle with a bundle containing a certain quantity of a single good of the first component. Think of this first component as income; it is not so unreasonable to say that our agent is indifferent between five dollars and five dollars worth of coffee.

Theorem 2

If the two agent preference ordering for $\precsim$ is transitive and complete while also satisfying monotonicity and income in bundle, then the case of intersubjective disagreement cannot occur. (Proof: *See Appendix.*)

We rule the above case out formally using the two properties for the sake of rigor. Without the two, our system can still make interpretations and is still useful. Furthermore, we could have simply, by assumption, limited our system to analyze preference orderings which do

---

[11] The axiom of strong monotonicity is often replaced by the weaker but more general axiom of local non-satiation. Depending on the context, the two can be equally general and some authors argue that strong monotonicity is implied by economic theory and need not come as an axiom. Below are some papers for the interested reader:
Becker, G. S. Economic Theory, Transaction Publishers, New Brunswick, N.J., 2008.
Border, KC. Lecture Notes: Monotonicity and Local Non-Satiation, April 2009.

not produce case 2. One may argue that the two properties are too restrictive or unreasonable, but they do not eliminate any real case of interest.

The first property is standard in economics because decisions are usually centered on "good" goods rather than bad "goods." Furthermore, one can make tweaks to compare preferences over bad goods while respecting this property.

The second property is named "income in bundle;" roughly it assumes the existence of a currency for which all goods can be traded. This need not be money; on a desert island, this might be food or whatever everyone finds valuable. One might argue that food is not currency, but on a desert island, food is potentially more tradeable; there are things people would not trade for money off the island that people would trade for food on a desert island. In short, this property denotes a "prime" good which exists to some extent in all societies. The extent of its tradeability is denoted by what other goods can be traded for it.

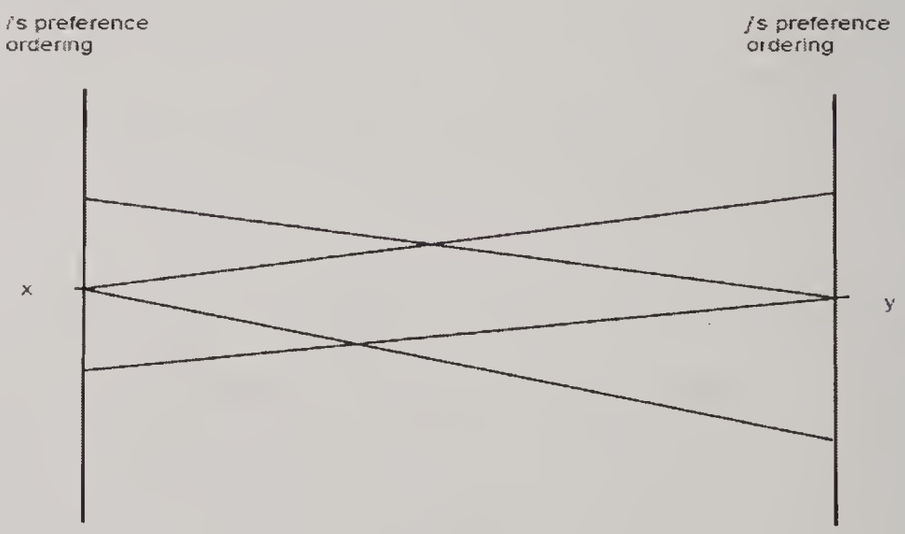



Fig.04 Similar but non-identical preferences (Case 3)

Case 3 is analogous to indifference. This is a case in which the agents cannot intersubjectively agree with each other. We need not rule it out as we did for Case 2 because there is no disagreement in a strict sense.

We now can move on to define the binary connectives of ICs.

The Binary Connectives

We use the same notation for the connectives as the single agent ones for sake of simplicity. It is clear which is which as the single agent ones have subscripts.

Weak Preference Difference

We say that agent *i*'s preference for bundle *x* is weakly stronger than agent *j*'s preference for bundle *y* if:

- *j* considers his preference for *y* to be weaker than *i*'s preference for *x*

*or*

- *i* considers his preference for *x* to be stronger than *j*'s preference for *y*

*or*

- *i* and *j* cannot come to agreement, as depicted in case 3.

We denote this by $[x]_i \gtrsim [y]_j$.

Strict Preference Difference

If the two first conditions are satisfied then it is a strict preference difference which is exactly the case of intersubjective agreement. Just like single person preference, the weak preference difference includes a possibility for the strict one.

We denote this by $[x]_i \succ [y]_j$

Incommensurable

We say that $[x]_i \sim [y]_j$ if neither agent believes the other's preferences are better satisfied.

Theorem 3

The order produced by $\lesssim$ is transitive and complete.

*Complete*: $[x]_i \gtrsim [y]_j$ or $[y]_j \gtrsim [x]_i$ must be true.

*Transitive*: if $[x]_i \gtrsim [y]_j$ and $[y]_j \gtrsim [w]_j$ then $[x]_i \gtrsim [w]_j$.

Proof:

-Completeness: See Appendix.

-Transitivity: Done by breaking the definitions into different cases which all satisfy transitivity.

Corollary

The three interpersonal relations have the same interpretation as the single agent ones if the two agents being compared are identical.

Namely $[x]_i \precsim [y]_i$ is equivalent to $x \precsim_i y$; the same holds for $\sim$ and $\prec$.

## IV. Social Choice and Welfare

In the following, I will define a social choice function (SCF) using the IC ordering obtained prior. I will then show that it has advantages over the utilitarian and egalitarian social choice functions.

For this section, we will be working with preference ordering which can be represented by continuous utility functions. An SCF is a decision rule. It is a means of choosing a social distribution given the different agents' preferences and a feasibility constraint. A utilitarian SCF is one which maximizes the sum of utilities of the agents; an egalitarian SCF is one that distributes the good equally amongst agents. They are defined formally as:

Utilitarian: $F(u_i,u_j,z) = (x,y)$ such that $\max\{u_i(y) + u_j(x) \mid x+y=z\}$
Egalitarian: $F(u_i,u_j,z) = (x,x)$ such that $2x=z$

The social choice I propose is the following:
$F(u_i, u_j, z) = \max\{u_i(x) + u_j(y) \mid x + y = z \textit{ and } [x]_i \sim [y]_j\}$

This implies that we maximize the sum of utilities, weighting agents equally, as long as neither believes the other agent is better off. We can show that this function is Pareto optimal and produces more total utility than the egalitarian choice, while obtaining a more equal distribution than the utilitarian function.

Theorem 4

The social choice function above is Pareto-Optimal. Namely, the constraint is non-binding with respect to optimality. (Proof: *See Appendix.*)

Theorem 5

The social choice function above always produces more or equal total utility than the egalitarian one. (Proof: Trivial, since the egalitarian distribution always satisfies the additional constraint.)

Theorem 6

The social choice function above always produces a distribution that is more or equally egalitarian than the utilitarian one. (Proof: Suppose that the two distributions are not the same. Then it must be that the utilitarian solution does not satisfy $[x]_i \sim [y]_j$. However, this is a constraint on how much allocations can differ, so it must be that the utilitarian solution is allocating goods in a less egalitarian manner. For some cases, $[x]_i \sim [y]_j$ is not an effective constraint, then the solutions will be the same for the two distributions.)

## V. Appendix

**Proof of Impossibility of Case 2 (by contradiction)**

1. Let agent *i* strictly prefer his indifference set for *x* over the interval equivalence set of *y* by agent *j*.
2. Let agent *j* strictly prefer his indifference set for *y* over the interval equivalence set of *x* by agent *i*.
3. By 1), there is a bundle $(a,0,...0) \sim_i x$ for all the *x* in $[x]_i$. By strong monotonicity, for all bundles of the form $(b,0,...0)$ in the interval set of *y* we have $a>b$.
4. By 2), there is a bundle $(b,0,...0) \sim_j y$ for all the *y* in $[y]_j$. By strong monotonicity, for all bundles of the form $(a,0,...0)$ in the interval set of *x* we have $b>a$.

**Proof of Completeness (by contradiction)**

Want to show: $[x]_i \succsim [y]_j$ or $[y]_j \succsim [x]_i$

1. Let $\sim([x]_i \succsim [y]_j)$ and $\sim([y]_j \succsim [x]_i)$
2. Then by $\sim([x]_i \succsim [y]_j)$:
    a) $j$ does not consider his preference for $y$ to be weaker than $i$'s preference for $x$
    b) $i$ does not consider his preference for $x$ to be stronger than $j$'s preference for $y$
    c) Not the case that indifference/Case 3 occurs.
3. And by $\sim([y]_j \succsim [x]_i)$:
    a) $j$ does not consider his preference for $y$ to be stronger than $i$'s preference for $x$
    b) $i$ does not consider his preference for $x$ to be weaker than $j$'s preference for $y$
    c) Not the case that indifference/Case 3 occurs.
4. We see that 2.a)b) and 3.a)b) give us exactly Case 3 which we have assumed would not occur.

**Sketch of Proof of Pareto Optimality**

The proof proceeds as follows:

1. I show that the constraint is equivalent to the indifference curves intersecting.
2. I show that the set of allocation where indifference curves intersect contains (0,0)
3. I show that the set of allocation where indifference curves intersect is unbounded.
4. I derive the Pareto Frontier and shows it intersects with the set.
5. Thus, there must be a Pareto solution within the constraint.

**Works Cited**

Arrow, Kenneth J. "Extended Sympathy and the Possibility of Social Choice." *Philosophia,* 7:233–237, 1978.

Harsanyi, John. "Cardinal Welfare, Individualistic Ethics, and the Interpersonal Comparison of Utility." *Journal of Political Economy,* 63:309–321, 1955.

Hausman, Daniel M. "The Impossibility of Interpersonal Utility Comparisons." *Mind.* 104:473–490, 1995.

Isbell, John R. "Absolute Games." In A. W. Tucker and R. D. Luce (eds), *Contribution to the Theory of Games, Vol. IV.* Princeton University Press, 357–96.

Robbins, Lionel "Inter-personal Comparisons of Utility: A Comment," *Economic Journal,* 48:635–641, 1938.

Rossi, Mauro. "Interpersonal Comparison of Utility: The Epistemological Problem." PhD Thesis, London School of Economics and Political Science, 2009.

I would like to extend my thanks to the helpful editors of Noēsis, in particular Mathew Armstrong and Howard Williams, whose work and feedback have greatly improved this paper. As well, Professor Peski for his invaluable help in the proofs of a theorem.

# Trust, Partiality, and Relationships: In conversation with Jon Rick

Mathew Armstrong, Jon Rick & Joy Shim

**Introduction:** Jon Rick has been instructing at the University of Toronto since 2013. In that time, he has been teaching a variety of courses, from seminars to surveys, in general ethics, climate ethics, ancient philosophy, and the relationship between interpersonal relationships and moral thought. A student favourite, Dr. Rick is known by many who have taken his classes to be an approachable and interested instructor who cares deeply about the learning of his students. Dr. Rick received all four of his philosophy degrees (B.A., M.A., M.Phil, and Ph.D.) from Columbia University, culminating in his dissertation, "From Partial Passions to Moral Sentiments: Taking Up Adam Smith's Impartial Spectator Perspective." On February 26th, 2016, we took the time to ask Dr. Rick a few questions about his research and teaching experiences, and Dr. Rick lived up to his reputation as an engaging, interesting philosopher.

**Noēsis:** We want to know how you first became interested in this discipline. What drew you towards studying and researching moral and political philosophy?

**Rick:** Well, when I was an undergraduate, I took a course (it was in my second year) in the history of Western political philosophy and moral thought—just a very wide-ranging course on classics. We read *The Republic* first and we'd move up to things in the 20th century, and I'd never read anything like that sort of stuff before in my high school years; it would've been mostly literature and textbooks. So, finding some writing that was argumentative and exploring questions about what the good life is and how we ought to live together was really exciting. These were questions that I, at the time, was trying to figure out myself, and it was a wonderful encounter with all these people saying really thoughtful,

insightful things about those matters, and it drew me in—I was excited. It was hard—I thought that this was something I could really use some help on, so I wanted to take courses in it and talk to professors and find out how to unlock the secrets. I was drawn in at that point and I continued to really appreciate these questions, which still seem urgent to me. It was really that course that drew me in, and it was fun.

I had a nice experience too where I later got to teach that same course, and I was very excited to have the opportunity to experience being on both sides of the table. For me, it certainly helped a lot having been a student in that setting to then teach that course. I always try to remember that—being on the other side of the table when I teach now. Hopefully I can bring some thoughts to class that help people with things they haven't thought of. And yet, it's certainly the case that I'm still learning. Still a student, just maybe one with a few more years behind me and a couple more pages turned. But I'm trying to work through these works and texts as well, and I think it's important for me to bear that in mind when I'm approaching a classroom setting. That idea of wanting to have some help with it—I always try to extend that back to my students. What can I do to make this text or argument engaging and open it up to people? When I'm reading, I can see where the text gets hard—there's a turn in the argument, there's something in the phrasing—plus, you can always trust a philosopher to use terminology in a particular way that can trick the students. So I come out there and say, "There's a trap! You have to watch out for this."

**Noēsis:** What are your current research interests?

**Rick:** I'm interested in topics in moral theory, particularly stuff that I've been doing a seminar on here for the past couple of years about relationships and about the moral standing of our relationships and our commitments to different projects as well. So these fit in the broad domain of topics about the moral intersection between impartiality and partiality. So there is stuff in that.

I also always have ongoing interests in the history of moral and political thought. I've been doing some work on Adam Smith's moral theory and some of his political economic writing. Some of that work has to do with intersections between Smith and Rousseau in moral psychology —about what has been called the drive for recognition—concerns about

having standing in the eyes of others and the role that plays in the moral psychology views of these 18th century thinkers. Also, I've been recently doing a little bit of work on Smith's political economic thinking with respect to learning some lessons, or trying to draw some lessons, about conviction and open-mindedness from what Smith has to say about financial crises in the 18th century and from how he changed his views on some of these matters. It seemed like a timely topic after everything that's unfolded and continues to be unfolding from 2008 onwards. It happens to be in the midst of some otherwise slightly dry writing on the history of money in *The Wealth of Nations*. There's some really wonderful stuff about financial regulation that I think is unexpected for many people who only know Smith from pundits and so forth. They treat him incorrectly as an arch-*laissez-faire*ist, a deregulatory fanatic, which he's not.

**Noēsis:** That's a very relevant connection between Smith and the modern financial situation. Adam Smith's reputation as that arch-*laissez-faire*ist does suggest an oversimplification of his position, but it is very interesting to hear about your reading of his complexity. In terms of partiality and impartiality in relationships, what do you think should be the role of special commitments in our ethical decision-making? Or, should we be more like impartial observers in our moral decisions?

**Rick:** I think that there ought to be a place within one's moral thinking for special obligations, special duties, and also just moral reasons to act in ways that express partiality, that treat some people differently than others. I think that treating people differently is essential to having intimate relationships from friendships to loves to families, and those are extremely valuable components of a good life. And if it is the case that they do require partial consideration, then that speaks pretty highly on behalf of partiality.

To borrow a turn of phrase from Bernard Williams, "If a moral theory said you couldn't have those kinds of valuable relationships, then you might say, so much the worse for the moral theory." At the same time, I think that there's a way in which you can think of a moral theory that incorporates rules or norms or reasons to treat people with partiality without necessarily having to rule out considering it to be an impartial moral theory. It depends on the level or order at which you require impartiality. It may be sufficient for many to say that as long as we're according the same rules or norms or reasons to others as we do to

ourselves, that we are achieving impartiality here. We're not making exceptions for ourselves or those we care about, say, from being bound by certain rules that we require of other people or vice versa. This is a kind of malignant or negative form or partiality that you'd want to avoid. But a theory that says it's permissible if not required to spend more time with your child than the neighbour's would be, to put it lightly, a benign form of partiality, to say the least—and it is one that is consistent with the sort of formal impartiality I've been suggesting. So I think there are certain theories that can have it both ways.

**Noēsis:** The example you gave for partiality, where you said it's okay for one to spend more time with their child than the neighbour's did not seem particularly morally charged. Could you perhaps give a more clearly 'morally charged' example illustrating the ways in which partiality can be used in ethics?

**Rick:** Well, I think that one's role in one's family has a lot of moral ramifications, so I'm not sure what you think by 'morally charged'; what do you have in mind?

**Noēsis:** I suppose that spending one's time could be seen just as attention, rather than something morally charged; perhaps it is just a desire that you might want to spend more time with your own child than another's, but then again, generally we think questions involving charity are morally charged. Money and time doing charity work can be examples of morally charged instances of using your resources, time, and money in a moral setting.

**Rick:** Let's take on Mat's notion of resources to explore some moral dimensions of family. Maybe this falls more into political philosophy, but we could ask what kind of inheritance laws we ought to have if a society is to be just. Maybe some would use those considerations to support allowing more freedom in the distribution of your resources. Of course, there would also be questions on how redistributive effects in your own society would work for establishing more relative equality and also for sending resources abroad, outside of your society, in the sense of charity Mat was alluding to. So often in these sorts of debates there are classic examples of treating everybody impartially; you used the phrase 'impartial

observer'—this is often taken to represent some detached, disinterested, austere individual who says we need to help people who are most disadvantaged; a representative of the view that there is no magic in the pronoun 'my' and that we must treat everyone equally.

I was suggesting that a theory allowing for differential treatment in that sense of partiality could avoid this type of substantive impartiality. To put it as a slogan, instead of treating everyone equally, we should treat them *as* equals, and maybe that's sufficient for getting all the impartiality you need in a moral theory. A more substantive impartiality across the board would run the risk of precluding or excluding these valuable relationships that do require differential considerations to those you care about as opposed to strangers or those you don't know.

**Noēsis:** That is an interesting balance of partiality and impartiality. Related to interpersonal relations, what role do you think the concept of 'trust' plays in moral philosophy?

**Rick:** This is a big question! Trust certainly plays a role in many theories of promise-making and promise-keeping, so you could think of trust as playing a central role in social contract thought. I also think it plays a role in the ethical considerations having to do with testimony and authority. There's a lot.

How does it relate to relationships? I think trust is very important—there are different ways in which you can talk about trust. One is more descriptive, one is a little more normatively charged. And these aren't the only ones. I'm thinking of this: one the one hand, you can trust that somebody will do something and have a straightforward descriptive expectation that they will do it (that these things will happen in the future). Then there is trust where it is some sort of moral attitude you take to be fitting or warranted for certain situations, where you have not just a descriptive expectation but a normative expectation to which you are holding others. And in all of our moral relations from the broadest to the most intimate, we have normative expectations that define the contours of those relationships, such that if they were violated, you might think this grounds for suspending the relationship. Classic examples of that would involve fidelity—a form of trust - in certain types of relationships.

I guess in that sense, I am fascinated with trust in relationships. I've been thinking of it recently in a slightly different context where I've

been wondering about how we trust some of our authority figures, certainly our political authority figures—thinking about how one can trust that they will follow through on a certain type of commitment that they are supposed to stand for. I am also very fascinated with how many relationships, like say a friendship, seem to demand that we ought to trust what our friends say and report more than that of a stranger, where it may not be the case that they have better access to the truth in these matters, but the relationship would erode in a potentially debilitating way if one didn't trust the reports of one's friends more than strangers. And this is one of the ways in which relationships and partiality can seem like a fun, puzzling topic. This is more like what we were talking about earlier. This relates to the way relationships can potentially problematize moral theories. In this question of trust you might think relationships are problematizing something like epistemic rationality, that there are certain norms we are supposed to follow if we're aiming at the truth or something like that. Yet, maybe relationships like friendships place certain requirements on our beliefs that vary from what we take to be standard norms of epistemic rationality. And that's a puzzle. For some, it might be easier to toss away a moral theory than it is to toss away truth-seeking rules and norms. So that's another place where relationships are puzzling.

**Noēsis:** That is interesting to see trust working in a way that is dependent on our relationships, rather than the other way around. Shifting gears from that, we would like to know what you think of climate change. Particularly, do you think we are morally beholden to future but as-yet unborn generations who will suffer most the effects of climate change?

**Rick:** You're really asking all the hard ones… not pulling any punches! What do I think about climate change? I think it's happening—anthropocentric climate change for sure.

But seriously, this is a difficult question. Maybe I'll back up for a second and say the question of whether or not we have moral obligations to future generations is a tough one because understanding obligations themselves involves many vexing questions. So, I'm going to back away from that and ask, do we have moral reasons to do things that would—hopefully the horse has not left the barn—mitigate climate change in a way that is intended to lead to a future that would have less suffering. My answer is yes. But I take it you're coming at this with some appeal (in the

background) to the non-identity problem, which is one of the more challenging questions in ethical thinking. Is it the case that if your actions confer existence on future persons, even if it's an existence with suffering yet one still worth living, you have done anything wrong? Could you say you've harmed them? It's one of these questions that people talking about climate change talk about a quite bit. People have different views on how to answer that question.

I think one of the things that is both great and incredibly vexing about talking about climate ethics is that it exposes just how puzzling certain ethical dilemmas can be when we work in an intergenerational context. Last term I did a course that used climate change to explore the non-identity problem, as well as another very difficult problem collective harm that occurs in some part because of this intergenerational component of climate change. Our traditional, common sense ways of thinking in moral theory, maintain that when wrongs occur, there will be an identifiable victim, and an identifiable perpetrator that exists at the same time. If you steal my car, I know who's at fault here. Assessing fault becomes difficult when we start talking about people who don't yet exist. Assessing blame is also difficult in the collective harm cases where the collective problem is too much atmospheric greenhouse gas. It seems clear that each of us is contributing to this problem, emitting some, but very little, comparatively, not enough to make a difference, it seems. Yet, all of us emitting together is causing a serious problem. But who is at fault, who is responsible – that becomes difficult. There is no identifiable perpetrator like when you stole my car (which I want back, by the way).

Those two, I think, are super puzzling questions, and climate offers a great opportunity to ask both. Regarding your question of how do we address our moral requirements or just the moral reasons we have to act in certain ways that will benefit future people, I think we do have reasons to leave things better than we're leaving them at the moment. But the question of whether we're obligated to so is hard as is the question of whether or not our reasons for leaving things better are grounded in not harming future persons. Moreover, it's genuinely tough when you're being asked to make potential sacrifices in your life that you won't see the benefits of. The ramifications or consequences won't be felt for many generations. So you can see why, for a lot of people, there is a lot of friction against doing anything. I do think that we have reasons to make the future

better for the people themselves, and we might have reasons to do it for the earth itself or other species as well. This gets into broad territory.

**Noēsis:** That seems to be the way it goes in Philosophy, and in ethics specifically; you think you are looking at one question and then find yourself looking at them all.

**Rick:** Yeah, that's right; there are a lot of rabbit holes to go down.

**Noēsis:** Thank you for going down some of them with us!

# Reference and Revision: In conversation with Imogen Dickie

Mathew Armstrong, Imogen Dickie, Joy Shim & Christopher Yuen

**Introduction:** Imogen Dickie began teaching at the University of Toronto in 2004. She is currently the Director of Undergraduate Studies in the St. George Philosophy Department. Her current research involves the theory of reference, singular thought, and associated topics in the philosophy of language, philosophy of mind, epistemology, and the philosophy of action. In addition to her teaching and administrative duties, she is an editor of *The Philosophers' Imprint*, and on the editorial board of *Ergo—an Open Access Journal of Philosophy*. Recently, she published a book, entitled *Fixing Reference*, developing an account of aboutness-fixing for thoughts about ordinary objects. We decided to interview her to hear some of her reflections on the publishing process and to gain some insight into her views on reference.

**Noēsis:** How did you first get into philosophy?

**Dickie:** I went to university to do physics and mathematics. Then in my first year, I had some extra space, and did a metaphysics course called "God, Mind, and Freedom". After that, I just drifted over to philosophy, partly because I was really terrible at making experiments work. That's actually very common; a lot of my colleagues started in the sciences.

Further back than that, I remember being shown Hooke's Law, which says how far a spring will stretch when you hang a weight from it, and thinking, 'That's not a law of nature!' This was about halfway through high school, so we were also doing Newtonian mechanics, and it seemed quite plausible that 'Force = Mass × Acceleration' was, somehow, a law. But the spring-stretching thing just looked like a statement of proportionality. In hindsight, the question of whether there really is a

difference in kind between laws of nature and mere statements of proportionality is very recognisable as a philosophical question. So perhaps that was an early sign that my future in physics was doomed.

**Noēsis:** What made you go into philosophy of language and mind instead of philosophy of science?

**Dickie:** I did my undergraduate degree in New Zealand, where I'm from. The university had a very, very small department, which did not offer a large range of courses. I did quite a lot of logic, and a lot of history of philosophy. When I applied to graduate school I said I wanted to do Early Modern—Locke, Berkeley, Hume, Descartes, Spinoza, Leibniz, and their friends. But that was really just because that was what I'd done most of in my first degree.

Quite early on at graduate school, I came across the questions about reference that I've been interested in ever since. From the philosophy of language side, the central question is what makes it the case that a singular term—for example, 'Bertrand Russell' or 'that'—stands for a particular object. From the philosophy of mind side, the central question concerns the aboutness of thoughts—how does understanding of a proper name or uptake from a perceptual link enable thought about an appropriately related thing? I became interested in these questions, read some very interesting work on them, enjoyed talking about them, and found that the people I wanted to work with were interested in them too. So having come in wanting to work on Descartes and Leibniz, I ended up writing a dissertation on direct reference.

**Noēsis:** It seems that there the influence on your work is coming from your teachers and the ones with whom you wanted to work. Do you find that the influence on your work sometimes goes the other way, that you have interactions with your students that influence what you're working on?

**Dickie:** Yes, very much so. My supervisor—John Campbell, who is now at UC Berkeley—used to say that if what you've written is not understandable by a pretty good second-year student, then you haven't made yourself clear enough. I think he was right about that. I think you should try to write for people who have done a bit of philosophy, but are not familiar with the topics you're discussing. By the time people are specialists in a fourth-year seminar, they already know a lot of philosophy, and aiming for readability by those people is setting the bar too low – it's making it too easy for yourself by supposing too much expertise in your

audience. So when I'm writing, I'm always thinking about my pretty good second year students, and whether they would be able to follow what I've written. If I think they wouldn't, I know I have to try harder. That's one respect in which interaction with students impacts on my research. I think if you weren't teaching then you'd just lose sight of what the standard for clarity is in philosophical writing.

As far as the content is concerned, quite often towards the end of a graduate seminar I'll have a couple of sessions about things I am working on myself, maybe even presenting my own stuff if I've got something that's far enough along for people to read. And to a lesser extent I'll do the same in a fourth year class if the students are interested. So I'll have students' questions and comments on the material. And I always take people's questions and comments very seriously. I keep minutes of feedback I get—if I go somewhere to give a talk, then before I get home I'll write out what happened in the discussion. I treat feedback from students the same way; it goes in the minutes, and I think about it as I'm trying to push my research forward.

**Noēsis:** Since we are talking about influences from others, students and teachers, on your work, you are also an editor on a couple of philosophy journals [*The Philosophers' Imprint* and *Ergo*]; we were wondering if that work, editing other people's papers, feeds back into your own work.

**Dickie:** Well, let me take the opportunity to say a little about these terrific journals, neither of which was set up by me.

*Ergo* is an open access journal that was set up by two of my colleagues, Jonathan Weisberg, who teaches at UTM, and Franz Huber, who is a St. George faculty member. It has a fantastic editorial model. Franz and Jonathan are the managing editors, so they're in charge, and are doing a lot of work. Then there is an *army* of section editors. I'm one of five section editors for the philosophy of mind. There is a similar little team for pretty much every area of philosophy you might think of.

When a paper comes to *Ergo*, if it's a paper in the philosophy of mind, then it'll to come to me or one of the other mind editors. Our initial job is to read the paper, decide whether to send it to referees, and if it is going to go to referees, find two people to do the refereeing. The 'army of section editors' model is a great thing about *Ergo* compared with many other journals. It means that every submission is read by someone in its field, whereas for most journals, you don't know if the person who's doing the first read and assigning referees is anywhere near your area. The 'army' model also means that each section editor has an agreement with

the journal to be responsible for around one paper a month. So if you are an author submitting to *Ergo*, you know your paper is not getting its first read from someone with twelve papers in front of them that they have to make a decision on that day.

If a paper does go out to referees, they'll send in reports, and the section editor will make a recommendation about publication. The whole process is triple blind. The author doesn't know who the editor or referees are. The editor doesn't know who the author is. And the referees don't know who the author is. It's as corruption-proof as I think it's possible for a journal's model to be.

*Ergo* is a new journal, though if you look at the papers it has published, you'll see that it's doing extremely well. The *Imprint* has been around a lot longer. It was founded fifteen years ago by David Velleman and Steven Darwall, who at the time were both at the University of Michigan. They were pioneers in bringing the ideals of open access publishing into philosophy—they just went ahead and set up from a standing start. Velleman designed the software and everything. I have a much larger role with the *Imprint* than with *Ergo*. At *Ergo*, there's the army of section editors, while at the *Imprint* there are six of us doing all the editorial work: Darwall, who's now at Yale, Velleman who's at NYU, Nishi Shah who is at Amherst College, Thomas Hofweber who's at Chapel Hill, Ian Rumfitt who is at Oxford, and me. The *Imprint* also has an excellent editorial model. If you send something to the *Imprint*, two of us will read it, and if we both agree that it should go out, then it will go out; if we both agree that it shouldn't go out, then it won't go out; and if there's a disagreement, then someone else will read it to provide a tie break. Again, everything is blind. And the editors are all puritanical anti-corruption fetishists, so we hope we are running a clean operation.

As I've said, these are both open access journals—as the *Imprint's* catchy jingle has it, 'Edited by philosophers. Published by librarians. Free to readers of the web.' One reason I agreed to be involved with them is that people I find it hard to say 'No' to asked me. But another is that open access journals are pushing back against the very questionable money-making model that many traditional journals are tangled up with. Philosophers who work in universities are all getting paid salaries, partly to produce the research that they submit to journals. And all the editorial and refereeing work that goes into a philosophy journal is being done by people from this same community of philosophers employed by universities. The journal publisher provides copy-editing and type-setting (which are often shockingly done, by the way), then sells the journal back to the universities for a profit, even though the universities have already,

by paying the salaries of the people who wrote the papers and volunteered their time as editors and referees, paid most of the real cost of its production! There's something pretty wrong with that. Many people think there should be a better way. Some of them—first David and Steve with the *Imprint*, now Franz and Jonathan with *Ergo*—have actually bothered to do something about it.

But all this has not yet answered Chris' question, which was about editing and how that feeds back into my own work. What we're looking for at both journals is an original contribution. Now, there's a lot of pressure on people to publish in philosophy. These days you probably need to publish to get a job (I was hired without being published, but this is increasingly less common); you need to publish to get tenure for sure; and then you just need to keep publishing because every year you've got to write a list of what you've done to hand in to the university, and if you've done nothing, you're conspicuous by the nothing that you've done. So people are under pressure to publish, and a result of the pressure to publish is that some not very good stuff ends up doing the rounds and eventually coming out. At the *Imprint*, David and Steve had the notion of what they called an "Intervention". An "intervention" is a paper which says "Philosopher A says X and Philosopher B says Y, and these are apparently inconsistent views, but look, if you slightly tweak A's position and you slightly tweak B's position, you find XY, a compromise"—something like that. Any paper which makes this kind of tiny move, we don't send out to referees. And I apply the same standard as a section editor at *Ergo*. So as an editor I'm looking for papers that make original contributions on genuinely philosophical questions. And, well, obviously that's the kind of paper I try to write myself, but, as one of my favourite colleagues likes to say, philosophy is hard....

**Noēsis:** So, on the topic of publications, you recently wrote a book; we just wanted to know what was the hardest part about writing the book, and how any of your views changed during the writing process.

**Dickie:** Do you want to see the book? [See figure 1.] The cover is by my friend who's a famous artist; she's a famous artist—she actually *is* a famous artist.

As for what's inside the cover, different parts were hard in different ways.

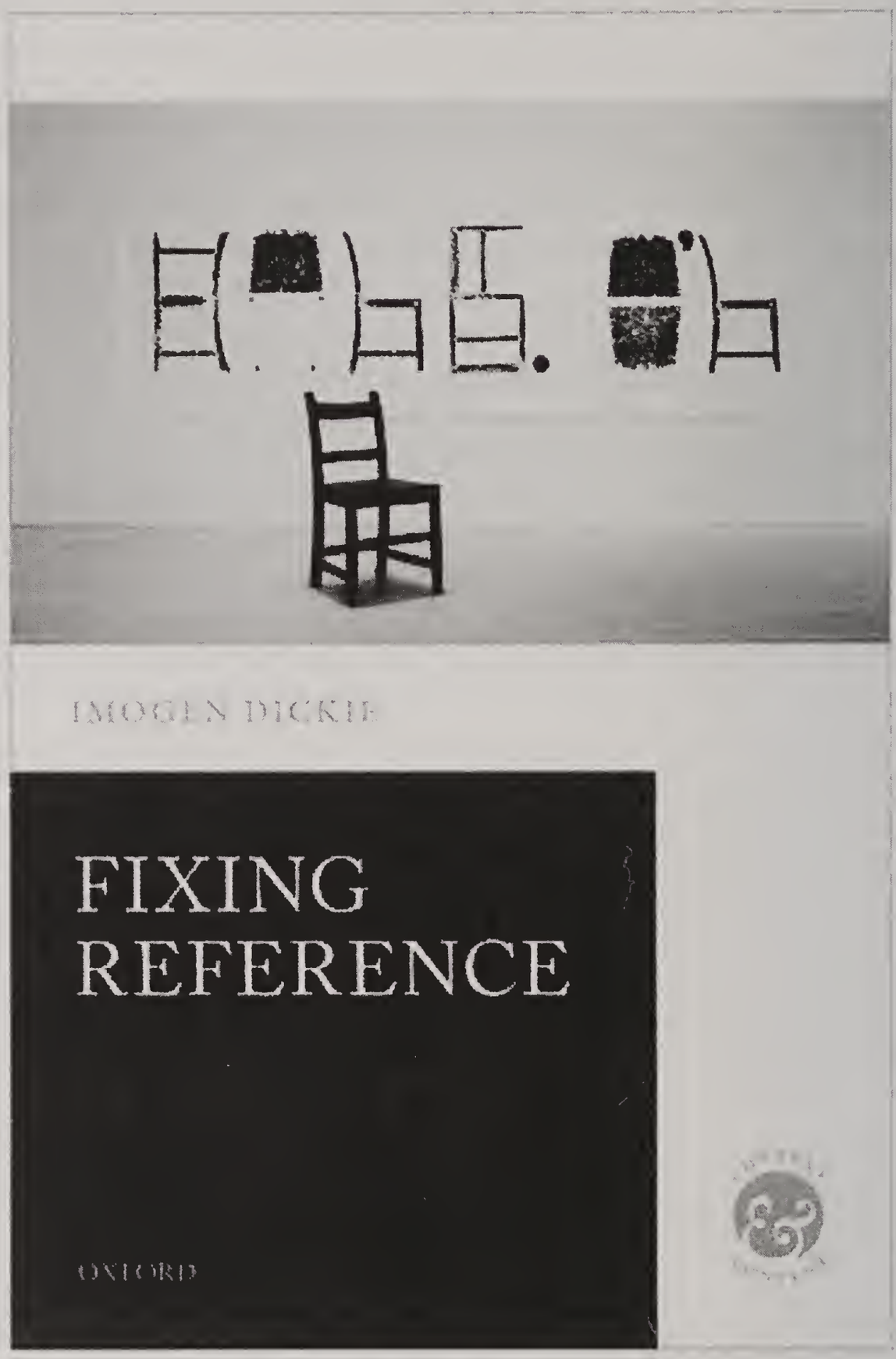



Figure 1. The cover of Professor Dickie's book, *Fixing Reference.*

The last phase was the production process, dealing with rounds of proofs, and that was really, extremely unpleasant, though I complained a lot to my more senior colleagues who said the production phase of a book is always horrible, and that made me feel a bit better. The book is published by Oxford University Press, but the production side—typesetting, and

putting the thing together—was all outsourced to a big international company. This company would send me a round of proofs. I'd go through it, find mistakes, and send back a corrected version. They'd send me a new round of proofs, in which the original mistakes may or may not have been corrected, but some innovator had introduced *new* mistakes. I'd go through the whole thing again, and send a new list of corrections. They'd send a new round of proofs, and so on. I had five rounds in all. So that was very frustrating.

Of course, before the production process there was the *writing the book* process, and that was hard in the characteristic way that philosophy is hard. There were some parts of the argument that took a long, long time to work out.

Something else that was hard was weighing up how much detail to go into against the need to avoid barriers to readability. It's asking quite a lot of a reader to expect them to get through pages and pages of philosophical text, so it's only fair to provide a narrative to give the reader a sense of moving along. I don't mean that I was trying to write a Harry Potter book, but I did want to write something that doesn't leave the reader feeling bogged down for pages at a time. But when you're trying to do that, and you get to an intricate piece of argument, you've got a problem: You need to decide how much detail to go into. On the one hand, there's no getting around the fact that the argument is intricate. On the other, you don't want to hold people up for twenty pages at the end of which you deliver one little sub-conclusion. To deal with this, something I ended up doing in a few places was trying to give a slightly thinned-down version of the argument in the main body of the chapter, then putting the details in an appendix. That way, if anybody actually shares my obsession about how the argument looks at the level of the tiniest little cogs and levers, they can look in the appendices.

So, in terms of the philosophical work (as opposed to the war of attrition with the production company) there were two difficult things: actually thinking of how stuff should go, and trying to write it up in a way that balances detail and readability. Actually, the whole thing was very hard. But it was a lot of fun, too.

**Noēsis:** What was the best part about dealing with the publisher?

**Dickie:** Peter Momtchiloff, the philosophy editor at OUP in the UK, has always been very good to me, and dealing with him is always a joy and delight – by the way, he's also a rock star (a real one with a band I mean); you can look him up on the internet. Peter knows the philosophical

profession very, very well—he knows who is interested in what, and so on. Anyway, he chose two anonymous readers to assess the manuscript I first sent to OUP, and they each wrote ten or twelve-page reports full of helpful comments and criticism. I don't know who these people are, except that they are Reader A and Reader B, but I feel like I love them. Their feedback was both incredibly encouraging and incredibly useful. This was actually the best experience with anonymous feedback I've ever had.

**Noēsis:** Was there specific feedback that you remember?

**Dickie:** It was just very interesting to see people's views on what worked and what didn't, and what was worth exploring more, and what they thought it would be good to stress in the final, polished version. A lot of the comments were pointed detail on specific parts of the argument. But they also had what I think was great advice on the strategic issue I was talking about before – the issue about balancing detail against readability.

Something that made the whole package of feedback really useful was that the readers seemed to be from very different backgrounds. The first reader was somebody who looked like they grew up in the same house I did—a probably-British-educated philosophy of mind/language person with interests not too remote from mine. The other reader seemed to be an American-educated, mind-metaphysics-y person. So their comments came from very different perspectives. And then the editor of the series the book is in, François Recanati, is a singular thought guy, which is very much the topic of the book, and he had a set of comments as well. So I had three sets of comments from different perspectives. I worked through every single comment, and made many changes, and I really can't thank these three people enough for their contributions to the finished product.

**Noēsis:** How long did the book take, from start to finish?

**Dickie:** A long time. I've been thinking about reference for years. My dissertation was on direct reference. There's none of the dissertation left in the book, except, right in the tiny details of the argument for the hardest step, a claim that I was obsessed with when I was a graduate student. This is in Appendix B of Chapter 2 of the book—my formerly favourite claim is a premiss in the argument! I was very pleased to find that. So the ghost of my dissertation is still there.

Anyway, I've been thinking about reference for a long time. But this particular project really began while I was at NYU from 2008–2010. At that time, I'd been mucking around with a view of what makes a

thought or sentence 'about' a particular thing for which I didn't have a first principles motivation. I had an example-driven motivation for the view. And I also thought it was the most plausible version of the view of reference in the *Tractatus*, by which I mean, what you get if you take the *Tractatus* picture and abstract away from the maniacal elements—if you do the abstraction you find a very powerful and intuitive proposal which didn't have a fair go in the 20th century because it got thrown out with the bathwater like the rest of the *Tractatus*. I'd had two papers published where I started to develop this kind of view. But I hadn't yet found the deepest formulation. In 2009, I was fiddling around thinking, "Where's the real motivation? Where's the real motivation?" —and then I found it! This book grew out of that. There's continuity with stuff I've done before, but this is really the working out of the framework that you get if you start with the first principles motivation that I found in 2009. So for the project that's six years, but it's not as though I started on it from cold. Philosophy just takes a long time.

**Noēsis:** Obviously the argument of the book is going to be complex, but is there a central argument that you can elaborate on for our readers that might make some sort of sense?

**Dickie:** Let me give it a try.

Think about views of reference-fixing that you might have had in an undergrad course. If you've had a course on this stuff, you'll have seen the contrast between descriptivist theories, where what makes it the case that the name refers to the object is that speakers associate the name with a description that the object satisfies, or maybe a cluster of descriptions, and causalist theories, where what makes it the case that the name refers to the object is that there's a causal chain leading back to the object (that's the picture in Kripke's *Naming and Necessity*). The late 20th century debate involved to-and-fro between these kinds of view. By now there's been a lot of work on this topic, but really no consensus. And I think that when you have a lot of ink spilled on a topic, over quite a long time, and no apparent progress towards a consensus, this a sign that people have been operating at the wrong level of explanatory depth.

So what I do in the book is go back to first principles. Suppose you've got a sentence, let's say 'a is F', and the question is what makes it the case that 'a' stands for a particular object. Causal theories and descriptive theories were attempts to answer this question. But I suggest that we step back (temporarily) from our obsession with what makes it the case that the name 'a' stands for object *o*, and reach for a principle at the

level of whole 'a is F' sentences or beliefs. The principle I reach for—are you ready? Here's a principle coming—is that justification is truth conducive. That's not to say that you can't have a false belief. It's to say that a factor that adds to the subject's justification for a belief somehow makes it more likely that the belief is going to be true, and that if you've got a justified belief which is not true, something's gone wrong. That's a principle connecting justification and truth. Now here's another principle, connecting truth and aboutness. If my belief that *Jack has fleas* is about my dog, then it's true if and only if he has fleas. So now we've got a principle connecting justification and truth, and a principle connecting truth and aboutness. [See Figure 2.] Let me say that again: *justification and truth*; *truth and aboutness*. And it's going to be just so disappointing if we can't cut out the intermediate term, and find a principle connecting justification and aboutness. This will be a principle which brings out the significance for accounts of the aboutness-fixing for beliefs, or reference-fixing for singular terms, of the fact that justification is truth conducive. The book argues for a precise version of the principle connecting justification and aboutness, and uses it to build an account of how aboutness-fixing works.

**Noēsis:** Can you tell us a bit more?

**Dickie:** Well, let's say a bit more about the principle that justification is truth conducive. One way to put this principle is as the claim that if you have a justified belief, you'll be unlucky if it's false, and not merely lucky if it's true. The principle connecting justification and aboutness that I propose can be put in terms close to this. Suppose you've got a body of beliefs which you'd express using a singular term like 'Jack' or 'Bertrand Russell' or 'that'. My principle says that these beliefs are about an object if and only if their means of justification converges on the object, so that you'll be unlucky if beliefs justified by this means do not match the object and not merely lucky if they do.

To get an intuitive handle on the view, think about what's involved when a telescope is focussed on an object. The fact that a telescope is focussed on an object doesn't guarantee that the data it delivers will match the object. But it does guarantee that if the data doesn't match the object, something's gone wrong – the situation is somehow unlucky. My framework treats aboutness as what I call "cognitive focus": A relation to an object puts you in a position to think about it by providing a means of justification such that, if you form only beliefs justified in this particular way but these beliefs don't match the object, something has

Figure 2. Professor Dickie's dog, Jack. We hope the sentence, *'Jack has fleas'* does not refer to him.

gone wrong—you're going to be unlucky if the beliefs don't match what the object is like and not merely lucky if they do. This isn't a descriptivist view, and it's not a causalist view. But it lets us explain when a description or causal relation is playing an aboutness-fixing role. A description or causal relation is playing an aboutness-fixing role if it's playing a role in securing justificatory convergence.

Anyway, that's just a start, but obviously this view of mine solves every extant problem about reference and aboutness!

**Noēsis:** How did your views change while you were writing the book?

**Dickie:** Well, I started out with an idea of how you would motivate a principle—this principle connecting aboutness and justification that we've been talking about. Then the hard graft in writing the book was in working out how you would actually argue for a precise version of the principle, and what its implications might be. I don't think that's really a case of changing your views. There are many, many things in the book which I could not have foretold were going to be there at the beginning of the project. But it was more a matter of bringing out what the view actually is, and what its consequences are for other philosophical issues.

In fact, a wise philosopher said to me once, when I told him I'd accidentally generated a solution to a problem that I wasn't actually trying to solve, that this is how you know you're onto something: If your view solves the problem that you set out to solve, well, don't be too pleased with yourself; but if it starts churning up solutions to other problems, then you know you've got something interesting. So often, especially when you go deep into a big, long project like this, you'll find surprises, things that you just didn't expect.

One big surprise was that the last chapter ended up being about the relationship between thought and consciousness. I had thought—and I thought this even until quite late in the game—that the last chapter of the book was going to be about something completely different. I thought it was going to be called "A Logical Atomist Revival Manifesto," and was going to be about my obsession with reviving logical atomism. But then I realized it had to be about something else altogether. Surprises just spring up in front of you as you're trying to bring out how your view actually works.

**Noēsis:** So rather than a process of changing your mind, it seems more like a process of discovering what's going on.

**Dickie:** Well that's right. You hope that it's genuine discovery and not just an illusion of discovery. So you hope that you're not just making stuff up! But what it feels like when you're doing philosophical work is that you've got the view in front of you on the table (as it were), and you're discovering how it works and what it entails.

**Noēsis:** Speaking of making stuff up, a lot of scientists might have a view against philosophy, I know a lot of scientists do turn into philosophers, but there does seem to be this perspective culminating in Stephen Hawking's declaration, "Philosophy is dead". What do you make of this sort of viewpoint?

**Dickie:** Well… now don't you think it's a bit premature, the declaration that philosophy is dead? I actually think people will always be interested in philosophical questions, and the real ones are questions that science probably can't answer. Of course, there was a time when philosophers were doing foundational work in the sciences—Descartes and Leibniz were major scientists as well as major philosophers. And those days are probably gone. But there are still philosophical questions: questions about what consciousness is; how the mind represents the world; the nature of proof; the difference between a genuine law of nature and something which is just a generalization; the difference between right and wrong; the nature of explanation; the relationship between morality and the law. These are questions of the kind that get people interested in philosophy in the first place. They're not going to be answered by the sciences. And as long as people are still interested in them, philosophy is not going to die.

And actually I think there are respects in which for some areas of philosophy scientific progress has brought us to a golden age. The questions that I'm thinking about in my book are one example of this. The amount of empirical scientific knowledge about our perceptual systems and about cognition that there is now enormous compared to what there was even a few decades ago, and I think this new scientific knowledge enables progress on some very old philosophical questions.

Let's take the case of 'that' beliefs formed on the basis of a perceptual link with the object. If you're going to have an account of what makes it the case that these beliefs are about the object, then you're going to need some kind of an account of how perception works. The early modern empiricists—Locke, Berkeley, Hume—all supposed some such account. But it was an account they were just making up! They said that your perceptual system sprays sense impressions or simple ideas at you, so that perception is giving you 'blue', 'square', 'fuzzy', then it's the job

of cognition to tie together the blueness, squareness, and fuzziness as properties of a single object. And this view of what perception delivers persisted well into the 20th Century—it's there in Russell and Quine, for example. But it's just empirically false. Psychology and neuroscience have given us a lot of new knowledge about perception and the boundary between perception and cognition. And this is a new tool to use in addressing philosophical questions about how perception enables thought. I think there's been a lot of recent progress on these questions made available by scientific advances.

Anyway, reports of philosophy's demise are greatly exaggerated.

**Noēsis:** That was an awesome answer. So rather than a murder-victim relationship you see science and philosophy in a kind of mutually beneficial relationship?

**Dickie:** Well look, I just talked about one respect in which scientific progress has enabled philosophical progress. You'd have to ask a scientist about the relationship going back the other way. And here I don't want to get above my pay grade by talking about the scientists' point of view. But it is worth noting that we don't have to go too far back to find major central figures in science who've had philosophical programs. Hilbert, a major mathematician, didn't like transfinite arithmetic, because he thought numbers you can get to by starting from zero and adding one are respectable, but numbers bigger than that, well… they're just spurious. So he wanted to try to find a way of using the tools that have been developed by people who use transfinite numbers, without being committed to the existence of such things.

Actually, maybe talking about Hilbert is going quite far back. But there are many people whose employers would classify them as scientists who are thinking about philosophical questions and who talk to philosophers all the time. The philosophical end of psychology is one obvious place where there are people like this—I'm thinking of Alison Gopnik and people like that (she's Canadian, by the way, and worked at the U of T early in her career). Constructivist mathematics is another example. What about physics, though? Well, there are quite a few people with backgrounds in both physics and philosophy thinking about things like the nature of space-time and the interpretation of the standard model of quantum mechanics. I went to graduate school with a number of people working on this kind of thing, and they're now employed in philosophy departments. Does Stephen Hawking think those research programs are dead? Or does he just say those people aren't philosophers because they

know some physics and are working on space-time and the interpretation of quantum mechanics?

**Noēsis:** Well we're almost out of time but before we finish, we just would like to know what new projects you're working on?

**Dickie:** At the moment I'm still on the clean-up operation from the book. I've got a bunch of papers I'm supposed to be writing which bring out connections between aspects of the book and other things. Aside from that, my new project is to find a new project. I worked towards the book for a long, long time, and I've really only just finished. And I can see little glimmerings sometimes on the horizon of what I might do next, but at the moment I'm still writing up unexpected solutions that my view's spat up.

**Noēsis:** No six months to pat yourself on the back and go, "Whew, six years? Okay, I'm going to be on a beach somewhere sipping martinis for a while"?

**Dickie:** Beaches and martinis is not my way of relaxing.

**Noēsis:** I suppose it'd be climbing, instead.

**Noēsis:** Do you find with this book, a lot of objections springing up that you will be responding to in papers or in other ways?

**Dickie:** The book only came out in North America a month ago, and in the UK just before Christmas, so it's still very early days. I hope that people will have things to say, and I'll respond when that happens.

**Noēsis:** Speaking of responses, in what form would you be responding to them? Do they just drop you an email saying this part doesn't really work and you reply to that, or is there going to be a second edition to the book at some point where you go, a year or two later, okay, these people have raised these issues and these are the revisions that have been made.

**Dickie:** We'll see, but there are a few things planned. There's going to be a symposium in the journal, *Philosophy and Phenomenological Research*; three people will read the book and write their reactions, which will be published with a response from me. There's going to be an author-meets-critics session at the American Philosophical Association in January. Again, three people will read the book and make comments, and I'll

respond, but this time it'll be a spoken thing. And sometime in the early summer I'll be the guest blogger on *The Brains* Blog—that means (apparently—this professor has never blogged before) that there will be a week where I do a series of posts about my book, and people comment on the posts or the book or whatever. As for dropping me emails—at the moment people are just writing to me and saying, "Hey I love the cover" [laughter] which is the one part of the book that I did not do. But people will write and have questions or links and connections with things that they've thought of, or there will be parts of it that they hate, and they'll tell me that, which I won't mind. The only bad outcome is if nobody reads it. That would make me sad.

**Noēsis:** Do you find a lot of the time when you get objections it's just from a misunderstanding of your view?

**Dickie:** If I get an objection from a misunderstanding, then my first instinct is to blame myself and try to explain myself more clearly. So hopefully this won't happen too much with the book, since so much work has gone into it. I've gone around giving papers on the material, and as I've said, I would write up minutes—I'd be there on the plane home, recording the facts that so-and-so said x and I said y and that didn't seem to work, so I said z and that seemed to work better, or so-and-so said something which has also been said a few months ago by so-and-so else. Then I'd try to improve my material in response to people's reactions and comments and criticisms and questions.

The point about misunderstanding is tied up with the thing I said before about writing strategy. One part of what you're trying to think about strategically while you write is the removal of barriers to readability. Another part is the minimization of the risk of misunderstanding. I mean it's a bit childish just to stamp your feet and say, "Oh everybody's misunderstanding me!" You've got to try very hard to make yourself clear. Of course, if you think about some historical philosophers there doesn't seem to be all that much effort in this direction. But you can't get away with that these days.

Quote on back cover: Simone de Beauvoir, *All Said and Done* (1972), p.16.
Original French: "Je me suis arraché dans le confort en toute sécurité des certitudes à travers mon amour pour la vérité et la vérité récompensé moi."

"I TORE MYSELF AWAY FROM
THE SAFE COMFORT OF CERTAINTIES
THROUGH MY LOVE FOR TRUTH
- AND TRUTH REWARDED ME."

-SIMONE DE BEAUVOIR